ETS TOEFL®

Monograph Series

MS - 29
April 2005

An Examination of Rater Orientations and Test-Taker Performance on English-for-Academic-Purposes Speaking Tasks

Annie Brown
Noriko Iwashita
Tim McNamara

An Examination of Rater Orientations and Test-Taker Performance on English-for-Academic-Purposes Speaking Tasks

Annie Brown

Language Testing Research Centre,
University of Melbourne, Australia

Noriko Iwashita

School of Languages and Comparative Cultural Studies,
University of Queensland, Australia

Tim McNamara

Language Testing Research Centre,
University of Melbourne, Australia

RR-05-05

ETS is an Equal Opportunity/Affirmative Action Employer.

As part of its educational and social mission and in fulfilling the organization's non-profit Charter and Bylaws, ETS has and continues to learn from and also to lead research that furthers educational and measurement research to advance quality and equity in education and assessment for all users of the organization's products and services.

To obtain more information about the TOEFL programs and services, use one of the following:

E-mail: toefl@ets.org
Web site: www.ets.org/toefl

Foreword

The TOEFL Monograph Series features commissioned papers and reports for TOEFL 2000 and other Test of English as a Foreign Language™ (TOEFL®) test development efforts. As part of the foundation for the development of the next generation TOEFL test, papers and research reports were commissioned from experts within the fields of measurement, language teaching, and testing through the TOEFL 2000 project. The resulting critical reviews, expert opinions, and research results have helped to inform TOEFL program development efforts with respect to test construct, test user needs, and test delivery. Opinions expressed in these papers are those of the authors and do not necessarily reflect the views or intentions of the TOEFL program.

These monographs are also of general scholarly interest, and the TOEFL program is pleased to make them available to colleagues in the fields of language teaching and testing and international student admissions in higher education.

The TOEFL 2000 project was a broad effort under which language testing at Educational Testing Service® (ETS®) would evolve into the 21st century. As a first step, the TOEFL program revised the Test of Spoken English™ (TSE®) and introduced a computer-based version of the TOEFL test. The revised TSE test, introduced in July 1995, is based on an underlying construct of communicative language ability and represents a process approach to test validation. The computer-based TOEFL test, introduced in 1998, took advantage of new forms of assessment and improved services made possible by computer-based testing, while also moving the program toward its longer-range goals, which included:

- the development of a conceptual framework that takes into account models of communicative competence
- a research program that informs and supports this emerging framework
- a better understanding of the kinds of information test users need and want from the TOEFL test
- a better understanding of the technological capabilities for delivery of TOEFL tests into the next century

Monographs 16 through 20 were the working papers that laid out the TOEFL 2000 conceptual frameworks with their accompanying research agendas. The initial framework document, Monograph 16, described the process by which the project was to move from identifying the test domain to building an empirically based interpretation of test scores. The subsequent framework documents, Monographs 17-20, extended the conceptual frameworks to the domains of reading, writing, listening, and speaking (both as independent and interdependent domains). These conceptual frameworks guided the research and prototyping studies described in subsequent monographs that resulted in the final test model. The culmination of the TOEFL 2000 project is the next generation TOEFL test that will be released in September 2005.

As TOEFL 2000 projects are completed, monographs and research reports will continue to be released and public review of project work invited.

TOEFL Program
Educational Testing Service

Abstract

This report documents two coordinated exploratory studies into the nature of oral English-for-academic-purposes (EAP) proficiency. Study I used verbal-report methodology to examine field experts' rating orientations, and Study II investigated the quality of test-taker discourse on two different Test of English as a Foreign Language™ (TOEFL®) task types (independent and integrated) at different levels of proficiency. Study I showed that, with no guidance, domain experts distinguished and described qualitatively different performances using a common set of criteria very similar to those included in draft rating scales developed for the tasks at ETS. Study II provided empirical support for the criteria applied by the judges. The findings indicate that raters take a range of performance features into account within each conceptual category and that holistic ratings are driven by all of the assessment categories rather than, as has been suggested in earlier studies, predominantly by grammar.

Key words: TOEFL, English for academic purposes (EAP), oral proficiency, second language acquisition, speaking tasks, integrated tasks

Acknowledgements

We gratefully acknowledge the assistance of a number of people who participated in this study:

- First, we are grateful to the 10 English-for-academic-purposes specialists who provided the verbal-report data that formed the basis of Study 1, the Rater Cognition study. These people were very generous with the time and dedication they devoted to the task we gave them.
- Second, we are grateful to the following graduate students who, as research assistants, played an integral role in carrying out the discourse analyses: Michael Fitzgerald, Jr., Felicity Gray, Daphne Huang, Clare Hutton, Debbie Loakes, Sally O'Hagan, Erich Round, Yvette Slaughter, Mary Stevens, and Susan Yi Xie.
- Particular thanks are due to Erich Round for his advice concerning the design of the phonetic analyses and to Sally O'Hagan for her invaluable assistance in the development of the analytic categories for the verbal-report analysis.
- Thanks are also due to Natalie Stephens for the immense amount of transcription that she undertook for this project; to Margaret Donald of the University of Melbourne Statistical Consulting Centre for statistical advice; to Paul Gruba of the Centre for Communication Skills and ESL, also at the University of Melbourne, for helpful feedback on drafts of the study; and to Debbie Wallace for dealing with the numerous administrative requirements of the project.
- Finally, we are grateful to Mary Enright of ETS for her assistance, advice, and encouragement in designing and completing this research and to the several anonymous reviewers of earlier versions of this report for their invaluable feedback on the work-in-progress.

Table of Contents

List of Tables

Page

List of Figures

Background

Introduction

Speaking tasks used in tests of English for academic purposes (EAP) increasingly seek to replicate the roles and demands of students in academic contexts. An important but rather underresearched development in EAP-test task design is that of *integrated* tasks (see Lewkowicz, 1997), in which test-takers are required to process and transform a cognitively complex stimulus (e.g., a written text or a lecture) and integrate information from this source into the speaking performance. Such spoken performances are more complex and more demanding than more traditional stand-alone or *independent* tasks, in which test-takers draw on their own knowledge or ideas to respond to a question or prompt. The absence of input in independent tasks means that these tasks are often restricted to fairly bland topics that draw on test-takers' general knowledge. Consequently, these conventional speaking tasks arguably underrepresent the construct of speaking within academic contexts.

Developments within the Test of English as a Foreign Language™ (TOEFL®) program (e.g., LanguEdge™) are beginning to explore the use of independent *and* integrated tasks to assess test-takers' readiness for academic study. (The rationale for the use of both independent and integrated tasks in the new TOEFL exam is articulated in some detail by Enright, Bridgeman, & Cline, 2002.) With regard to these developments, the need for valid criteria for the assessment of performance on complex integrated tasks emerges as urgent. Given that performance on such tasks involves the integration of cognitive skills (information selection and structuring) with more conventionally defined language proficiency skills, the question of the nature of the relationship between the two, and in particular how they may be judged together, is particularly vexed. To what extent do raters attend to each of these different dimensions of performance? And how do their assessment orientations on these types of tasks differ from those on tasks with limited input? Another question of importance involves whether and to what extent test-taker performance on integrated tasks differs from that on more conventional tasks. Does the heavier cognitive load alter the performance, for example, in ways other than the quality of the content?

The two related studies reported in this paper address these and other questions. The first, the rater cognition study, is an examination of what EAP specialists claim to value when they assess performance on independent and integrated tasks. The second, the speech samples study, is an analysis of test discourse pertaining to the same tasks. The two studies provide complementary

perspectives on the assessment of EAP speaking proficiency: One examines the nature of field experts' orientations and perceptions, while the other explores the empirical reality of the discourse. Together they can inform the development of empirically based, valid, and useful scales for the assessment of performance on independent and integrated speaking tasks (a topic that is discussed in more detail in the next section).

Study I utilizes verbal protocol methodology (Ericsson & Simon, 1993; Green, 1998) to investigate the unguided judgments of university-based oral-communication-skills specialists on conventional (independent) and more cognitively demanding (integrated) speaking tasks. Verbal protocols have been used for many years to explore the cognitive processes of *learners* of a second language. More recently, a number of studies have used the methodology to investigate the cognitive processes of *raters* of second-language writing performance, following the lead of studies in the field of first-language composition assessment. These studies have addressed questions such as how raters approach the task of making an assessment, how training or experience affects raters' behavior, and the features of performance on which raters focus. As yet, however, there is little research examining in detail what raters attend to when judging performance on speaking tasks. A handful of studies exists that examine raters' strategies on general-purpose second-language oral proficiency (these are described in a later section), but none involve EAP speaking tasks and none use integrated speaking tasks.

Study I is distinguished from previous research by its breadth of focus on rater cognition across a range of speaking tasks and, in particular, its concern with distinctions between independent and integrated tasks. Such comparative information has the potential to contribute to understandings of the demands made by such tasks and the ways in which performance might be judged. We ask the following primary questions:

- Are different criteria more or less salient in the different task types?
- Are different criteria operationalized differently in the different task types?
- How are increasing levels of proficiency characterized in the two task types?
- In what ways does the information raters heed in integrated tasks warrant the development of new types of rating scales?

Study II addresses the relationship between field experts' perceptions of test-taker

performance (the scale descriptors) and test performance itself by means of a detailed analysis of test-taker discourse on the same tasks. Focusing on performance features identified in the rater cognition study, the speech samples study aims to identify those features of performance that distinguish levels of proficiency. It also explores whether differences in task performance predicted by task specifications and judges' perceptions are empirically measurable or observable within the discourse produced by test-takers.

The two studies complement each other in that the empirical analysis of test-taker discourse can provide empirical evidence of the validity of judges' perceptions of proficiency. In addition, the evidence, on the one hand, of domain specialists' ways of conceptualizing proficiency and, on the other, of empirically observable features of test-taker discourse can together contribute to the empirical development of contextually valid new scales by ensuring that the content of the scales is relevant to performance on the particular tasks *and* reflects the features deemed by expert judges to be relevant to their assessments of proficiency.

The Empirical Development of Rating Scales

One of the questions addressed by this two-part study concerns the value of verbal-report data generated by field experts for the validation or empirical development of scales. It has been pointed out by a number of writers (e.g., Cumming, Kantor, & Powers, 2001, 2002; Fulcher, 1987; Matthews, 1990) that rating scales used in the assessment of second-language proficiency often have no basis in actual performance. North and Schneider (1998) comment that "[m]ost scales of language proficiency appear in fact to have been produced pragmatically by appeals to intuition, the local pedagogic culture and those scales to which the author had access" (p. 220). Citing critiques of scales such as the American Council on the Teaching of Foreign Languages Proficiency Guidelines (e.g., Bachman & Savignon, 1986; Lantolf & Frawley, 1988) and the Australian Second Language Proficiency Ratings (e.g., Brindley, 1986; Pienemann & Johnson, 1987), they argue that while an "intuitive approach" to scale development may be acceptable in low-stakes contexts, it is not appropriate in high-stakes contexts.

North and Schneider (1998) go on to review empirically based approaches to scale construction, citing the following approaches described in the literature:

1. the intuitive identification of key "features" at different levels through rater discussion of performance samples ranked in consensus order (Alderson 1991; Mullis, 1980)

2. the identification of decision points (Upshur & Turner, 1995)
3. scaling of key target behaviors using Rasch analysis (Griffin, 1990)
4. scaling of items using Rasch analysis (Brown, Elder, Lumley, McNamara, & McQueen, 1992)
5. discriminant analysis (Fulcher, 1993);

As North and Schneider point out, the first two approaches to scale development draw on expert judgment about test-taker performance. The first approach was usefully applied to second-language test data for the development of EAP writing scales for International English Language Testing System (IELTS; see, e.g., Alderson, 1991), and is the approach that is most closely aligned with that used in the first part of the current study. For development of the IELTS scales, after grouping IELTS writing scripts by level, EAP experts were asked to discuss them and agree on the key features of each script. Features that were characteristic of each level were then identified, and these in turn enabled the identification of criteria for assessing performance at different levels of achievement. These were then synthesized into a hierarchical set of descriptors for use with IELTS essay prompts.

The second approach, which was suggested by Upshur and Turner (1995), is relevant to the development of a primary-trait scale rather than to a generic one, and leads to scales that are somewhat different from traditional ones. In these "empirically-based, binary-choice, boundary-definition" scales, the final rating is not determined through the matching of a performance to one (of several) hierarchical scale descriptors that describes it most closely, but through a series of yes/no responses to questions about the quality of the performance.

The third approach—that employed by Griffin (1990) in the development of a set of scales for the assessment of literacy and numeracy skills in adults, the Australian Literacy and Numeracy scales—also drew on field experts as informants, namely literacy and numeracy teachers. However, the scaling of the descriptors was the end of a complex procedure that involved, first, the identification of key target behaviors abstracted from actual performance; second, the gathering of questionnaire data that reflected teachers' attempts to relate the key behaviors to their students; and third, the scaling of the behaviors using Rasch techniques to analyze the questionnaire data. However, this approach is based on teachers' intuitive evaluations of what their students can do rather than on actual performance.

One reservation that North and Schneider (1998) and Cumming et al. (2001) both raise in the use of expert judgment for scale development is that of agreement among raters as to what the key features are. This point is also made by Norris (2001) when he comments that the influence of individual participants must be considered in determining the eventual validity of inferences based on the resulting rating criteria. Perhaps in practical terms what this concern does imply is that it is necessary to involve a number of experts and to ensure that the resulting criteria or descriptors reflect the views of the majority of them, rather than perhaps one or two of the most vocal or influential. This is essentially an issue relating to the validity of the development procedure. In the model Norris describes for the development of criteria for EAP speaking and writing tasks, he deals with this through a thorough documentation of the responses raters generated as they reviewed test taker performances. However, ultimately the identification of the final task criteria is accomplished in much the same way Upshur and Turner (1995) and Alderson (1991) describe—namely, a process of consensus building.

Brindley (1991) raises another important issue in the use of field experts to develop rating criteria: Who should be viewed as an "expert?" As he points out, conventionally it is language teachers, although others possible experts might include test users, the learners themselves, and people with whom learners will interact in the target context (e.g., employers, lecturers) or "naïve" native speakers. In Norris's approach to the development of criteria for EAP writing tasks, he includes three groups of experts—ESL specialists, discipline-based university lecturers, and students who have had some experience of coping in the target context.

Of the other two approaches reported by North and Schneider (1998), the fourth relates specifically to the development of scales describing performance on discrete-point tests rather than productive performance, such as essay writing or speaking, and as such is not relevant to the present study. The fifth approach does not draw on experts but involves the analysis of actual performance. In the study conducted by Fulcher (1993, 1996), the aim was to identify fluency phenomena that reflect different levels of proficiency through the analysis of audio-recordings of oral interviews.[1] A number of fluency phenomena were identified, and there was a high degree of concurrent validity in that the features predicted scores to a high degree. However, the saliency of the phenomena identified by the analyst to the raters was not checked, thus little can be concluded about their validity as rating criteria in terms of whether raters actually attend to them.

The present study seeks to avoid the limitations of these earlier studies by taking a twofold

approach in which the analysis of verbal reports (i.e., descriptions of what judges attend to while rating performances) is complemented by a discourse analysis of the performances on which those evaluations are based. It is argued that such a study has the potential to inform the development or validation of rating scales; by providing evidence of the qualities of performances on which judges focus when they are not given explicit direction, such information can ensure that the content of the scales is *relevant* to the context, *meaningful* to raters, and, when complemented by an analysis of test discourse, *observable* in task performance.

Rater Cognition Studies in Oral Assessment

Although there is a growing body of literature on rater cognition in both first- and second-language writing contexts (e.g., Cumming et al., 2001, 2002; Huot, 1993; Milanovic & Saville, 1994; Milanovic, Saville, & Shen, 1996; Vaughan, 1991; Weigle, 1994), research into the cognitive processes employed in the rating of oral proficiency is extremely limited. Three studies that explore rater cognition in the assessment of second-language oral proficiency are relevant to the present study. Two of these, Meiron (1998) and Brown (2000), are concerned with the validation of an existing scale, and one, Pollitt and Murray (1996), with the elicitation of unguided orientations among experienced raters of second-language oral proficiency. Meiron's study used a tape-based monologic task, whereas the other two relied on dialogic performance in oral interviews.

Meiron (1998) explored rater behavior on a single SPEAK® task, the picture narrative, in which learners retell a story using a series of picture prompts. Two dimensions of behavior were explored, both the rating focus and the methodology raters adopted to carry out ratings. Meiron found that in addition to the specified criteria, raters also reported using certain self-generated features not mentioned in the scoring rubric. Two approaches to assessing performance were identified: a "quasi-analytic rating" in which discrete features in the speech sample were differentially weighted to arrive at a final score, and a "more truly 'global' or 'holistic' assessment" in which the rater did not focus on any one specific feature.

Like Meiron, Brown (2000) found that raters focused not only on criteria specified in the scales (i.e., syntax and vocabulary), but also on aspects of performance that were not explicitly included in the scales—namely, pronunciation and fluency. Raters also commented on aspects of "communicative skill," a loosely defined orientation within the scales which appeared to include a range of behaviors as broad-ranging as the use of communication strategies, test-takers'

comprehension of the interviewer, the (perceived) ability or willingness to take the initiative or talk at length, and aspects of discourse (its structure and organization as well as its content). In addition, as specified in the scales, raters also focused upon the fulfillment of the functional demands of the task (i.e., narration, description, giving opinion, and hypothesizing). However, Brown also found that raters differed in their views of what constituted fulfillment of the functional demands of the task: Some raters took a "narrow" approach, looking for the use of particular linguistic structures typically associated with certain functional demands (such as the comparative for comparison and the conditional for speculation), while others were more concerned with whether test-takers' responses were contextually appropriate in terms of content. An additional finding was that inferences were frequently drawn regarding test-takers' ability to cope with real-world (i.e., academic) demands, and these were based to a large extent on the content of test-takers' contributions and their interactional styles.

While neither Meiron (1998) nor Brown (2000) report finding any differences in assessment focus according to the level of proficiency, Pollitt and Murray (1996) found that different performance characteristics were more or less salient at different levels of proficiency. They found that when assessing performance on the Cambridge Assessment of Spoken English oral interview and not guided by a specific scale or set of criteria, raters focused more on grammatical competence at the lower levels and more on sociolinguistic and stylistic competence at the upper levels. They also found that comprehension was particularly problematic; there was some concern among the raters as to whether inappropriate responses should be penalized given that they frequently arose from comprehension problems and were therefore not necessarily part of oral production, which is what the test was intended to assess.

One limitation of verbal-report studies involving assessments of speaking proficiency rather than writing proficiency is that the real-time nature of the assessment precludes the elicitation of concurrent reports, and limits, therefore, what can be inferred about the *process* of rating, as opposed to the performance features to which raters attend. Whereas recent studies of the assessment of writing (e.g., Cumming et al., 2001, 2002; Lumley, 2000; Milanovic, Saville, & Shen, 1996) have identified a range of approaches to assessment in terms of the procedures raters follow in order to arrive at a judgment of proficiency, this study is concerned not with such mental processes, but simply with identifying the features on which raters focus when attempting to reach that judgment.

Discourse Studies in Oral Assessment

An increasing volume of research in language testing has analyzed test-taker discourse in oral assessment. Examining closely what test-takers produce can provide useful information in test validation. Shohamy (1994) argues that insights from discourse analysis provide a significant contribution to defining the construct of speaking in oral tests in general. Likewise, van Lier (1989) stresses the importance of discourse analysis, especially the need to look at oral tests from within and to analyze the test as a speech event in order to address issues of validity. McNamara, Hill, and May (2002) provide a recent survey of studies of oral test discourse.

The bulk of research on test-taker discourse analysis in oral assessment has been conducted in the context of the oral interview and has investigated ways in which features observed in oral proficiency interviews are different or similar to conversation (e.g., Johnson, 2000; Lazaraton, 1996; Young & He, 1998). Research has also examined variational features among interviewers and their potential impact on test-taker performance (e.g., Brown, 2003; Cafarella, 1997; Ross & Berwick, 1992). Other studies have adopted aspects of methodology used in interlanguage analysis in second-language acquisition studies, and have cross-referenced their findings to test scores in order to examine the potential effect of test-performance condition (e.g., Iwashita, McNamara, & Elder, 2001; Wigglesworth, 1997), test methods (e.g., O'Loughlin, 1997), interlocutor effects (e.g., Iwashita, 1996), and test-taker characteristics (e.g., O'Loughlin, 2002) on test performance.

Although the number of studies investigating test-taker discourse has been growing, to date few studies have examined the relationship between test score and the substance of the performance on which it is based in order to validate the rating scales used in an assessment. Two main groups of studies that have addressed this question have reached very different conclusions. Douglas and Selinker (1992, 1993) argued that, despite scoring rubrics, raters may well arrive at similar ratings for quite different reasons. In other words, speakers may produce qualitatively quite different performances and yet receive similar ratings. Building on studies by Douglas and Selinker (1992), Douglas (1994) compared test scores awarded to six graduate students from Czechoslovakia with transcripts of their semi-direct oral test performances. Various aspects of test-taker responses (e.g., local and global errors, risky versus conservative response strategies, style and precision of vocabulary, fluency, content, and rhetorical organization) were analyzed in order to compare the actual language produced by subjects who received similar scores on the test. The results revealed that very little relationship was found between the scores on the test and the

language actually produced by the subjects. Douglas speculated that one of the reasons for the discrepancy could be that aspects of the discourse that were not included in the rating scales influenced raters. It is generally accepted that language in use is a multicomponential phenomenon, and thus interlocutors' interpretations of a message may vary according to the facets to which they attend and how those features interact. Douglas suggested think-aloud studies of rating processes be undertaken in order to understand more thoroughly the bases upon which raters make their judgments, a strategy we have adopted in the present study.

In contrast, Fulcher (1996) is more optimistic about the relationship between characteristics of candidate speech and the wording of rating scales. He analyzed the transcripts of 21 English Language Testing Service (ELTS) interviews in terms of the rating category "fluency." Using Grounded Theory Methodology (Strauss & Corbin, 1994), eight different aspects of fluency were considered on the basis of examining the interview transcripts in detail. All 21 transcripts were coded into the eight explanatory categories and further cross-referenced with ELTS band scales using discriminant analysis. The results showed that all eight explanatory categories taken together discriminated well between students. The relationship between actual band scores and predicted band scores was further examined by comparing which bands would have been awarded purely on the basis of the explanatory categories. Only one case out of 21 was awarded a different band score from the actual band score.

Given these contrasting findings, the fact that little is known about the performance features to which raters actually attend, and how this relates to what test-takers produce, we decided to do a careful comparison of test-taker discourse and rater protocols in this study through the judicious selection of discourse analyses that reflect important categories and features identified by field experts.

Overview of Study I and Study II

The two current studies draw on a single set of test performances that were elicited using a set of independent and integrated speaking tasks developed for the new TOEFL project. In Study I, the rater cognition study, these performances formed the stimulus for raters' verbal reports; in Study II, the speech samples study, the same performances were analyzed using categories and features identified by raters in the rater cognition study.

Study I: The Rater Cognition Study

Verbal reports produced by expert judges (university-based ESL/oral communication skills specialists) while evaluating the quality of test performances form the basis of Study I. As the study was concerned with identifying appropriate criteria for the assessment of test performance, rather than determining how well raters were able to apply specified criteria, the judges were not explicitly guided as to what features of performance they should consider. Rather, the procedure was designed to elicit the understandings of university-based oral communication specialists as to constructs of oral EAP proficiency.

In order to explore the questions identified earlier, we framed the following research questions to guide Study I:

1. To which conceptual categories do "field expert" EAP judges attend in conducting their evaluations of performance on oral test tasks?
2. How do these judges characterize increasing levels of oral proficiency?
3. To what extent are these conceptual categories task- or task-type specific?

Findings are later presented in terms of these research questions. A typology of the conceptual categories to which judges attended in general across tasks and levels was chosen to address Question 1, while descriptions of increasing levels of performance for each of the conceptual categories, synthesized from the judges' comments, was selected to satisfy Question 2. An analysis of the task-specific nature of the assessments in relation to independent and integrated tasks was adopted to explore Question 3.

Study II: The Speech Samples Study

In Study II, the test-taker speech samples were analyzed using a variety of measures. The aim of the analysis was to examine the extent to which descriptions and evaluations produced by the judges were borne out in the reality of test-taker discourse—that is, to seek evidence that would support or challenge the validity of the judges' responses. For this reason, the measures used in the analysis were derived from the assessment categories identified by judges in the rater cognition study. More specifically, the speech samples analysis was also concerned with the extent to which differences in performance were observable or measurable across levels and across task types.

The following research questions guided Study II:

4. How can increasing levels of oral proficiency be characterized, and to what extent do the findings of the speech sample analysis match judges' perceptions of increasing levels of proficiency?
5. To what extent does an empirical analysis of test-taker discourse on different task types match differences in performance identified by judges and foreshadowed in the task specifications?

Study II findings are presented later as a comparison of discourse measures across five levels of proficiency (Question 4), a comparison of discourse measures across the different task types (Question 5), and a discussion of the relationship of the discourse findings to the outcomes of the rater cognition study.

Study I: The Rater Cognition Study

Methodology

Five tasks developed for the new TOEFL project were selected for use in the present study. As shown in Table 1, these included two independent speaking tasks, two integrated listening-speaking tasks (one based on monologic input and one based on dialogic input), and one reading-speaking task. Appendix A provides and describes the actual prompts.

Test data pertaining to the five tasks was collected by way of digital recording during the pilot testing of the tasks conducted by ETS® in the United States. All of the trial performances were double-rated by ETS staff using a purpose-designed draft scale with five levels. For the purposes of this project, eight samples at each of the five levels were selected from a larger pool of digitally recorded pilot-test data—a total of 40 performances per task and 200 performances altogether. While no information was provided on interrater reliability, the samples were chosen where possible on the basis that both ETS raters had awarded the same score.[2] The samples were randomly coded in order to obscure the proficiency levels assigned by ETS, then copied onto a CD-ROM.

Booklets were prepared in order to provide judges with information about the tasks used in the study. For the independent tasks, the task rubric and prompt were provided, along with information on the targeted functions and discourse features (see Table 1). For the integrated tasks, the input text (lecture script or reading passage), listening or reading comprehension tasks, and task rubric and prompt were provided, along with information on the targeted functions and discourse features (see Table 1). The task specifications were also provided in order to make clear

to the judges what skills the tasks were intended to measure. However, as the aim of the study was to investigate field experts' unguided orientations, no criteria or scales were provided.

Table 1

The Tasks

	Task	Type	Targeted functions and discourse features	Preparation time (s)	Speaking time (s)
1	12-month	Independent	Opinion; Impersonal focus; Factual/conceptual information	30	60
2	Art and music	Independent	Value/significance; Impersonal focus; Factual/conceptual information	30	60
3	Groundwater	Integrated; Monologic lecture	Explain/describe/recount; Example/event; Cause/effect	60	90
4	Rhesus	Integrated; Dialogic lecture	Explain/describe/recount; Process/procedure; Purpose/results	60	90
5	Innate	Integrated; Reading	Explain/describe/recount; Process/procedure; Purpose/results	90	90

Ten EAP specialists with experience teaching oral EAP skills in a university setting were recruited to take part in the rater cognition study. All were qualified teachers with a degree and postgraduate teacher training, or other training, such as the Royal Society of the Arts certificate or diploma. Most possessed postgraduate research qualifications as teachers of English to speakers of other languages or in applied linguistics. Of the 10 participants, all had experience teaching oral communication skills to nonnative speakers of English, and some has also taught oral communication skills to native speakers. The judges were given initial training and practice in verbal-report production.

Each of the judges provided verbal reports for one performance at each level on each task—a total for each judge of five verbal reports per task and 25 in total. The samples were allocated randomly, and in random order, and judges were not informed of the scores that the performances had been awarded. The data gathering was spread over two sessions of 3-4 hours

each. The first session began with familiarization and training in the production of verbal reports.

For each performance, the verbal-report procedure consisted of two distinct but related activities. First, the judges were asked to listen to the performance as they would when rating (i.e., essentially straight through, but with repetition where needed) and then to describe their overall impression, using any terms with which they felt comfortable. Second, they were next asked to elaborate on their overall evaluation by pointing out specific features of the performance that affected their overall assessment. To do this, they replayed the performance, stopping the tape at intervals to comment on specific features.

Analysis

The verbal-report data were transcribed and checked. Each transcribed report was then divided into naturally occurring speech units, which included the overall assessment of the performance (the summary turn) and various statements about specific features of the performance (a number of review turns). These turns were naturally occurring in that they were separated by replayed sections of the test-taker performance. Procedural comments such as "I'll go on now" or "That's all for that one" were removed, and comments which were repeated in consecutive turns were conflated into one. Finally, each turn was entered into a database for further segmentation and coding. (Appendix B provides a sample verbal report.)

The first step in analyzing the transcribed data was to scan the reports to identify rating orientations and develop an approach to segmentation and coding. Using a draft set of categories based on those identified in a pilot study (Brown, McNamara, Iwashita, & O'Hagan, 2001), two coders undertook repeated attempts both together and independently to segment the data into "ideas" units and to code them. An "ideas" unit is defined by Green (1998) as "a single or several utterances with a single aspect of the event as the focus."

It was determined that five types of ideas units could be easily and reliably identified, each concerned with a distinct aspect of performance and together accounting for almost all of the data. Four of these reflected the types of conceptual categories often used to define analytic criteria on assessments of speaking—namely linguistic resources, content, phonology, and fluency. The fifth consisted of overall or global assessments. The conceptual categories themselves (including the justification for this particular division and, in particular, the categorical differences in this and the previous study) are described in the results section of this report.

On the basis of the initial analysis, the principal researcher developed a draft coding

protocol that described the categories and dealt in detail with the most problematic coding decisions. Once this had been done, the analysis proceeded iteratively. It consisted of two repeated phases—coding and checking. In the first phase, the two coders worked independently to identify and code ideas units within a subset of the data according to the protocol; this was followed by checking for agreement and, where necessary, amendment of the category descriptions. The two steps were repeated, sampling from all judges, tasks, and levels, until a high level of consistency was achieved (0.9 or higher was considered good). After this, the two coders independently coded the entire data set.

The segmentation and coding of the review turns was relatively straightforward because they were usually short and were concerned with a specific extract from the performance, which tended to consist of only one or two ideas units. The summary turns, however—which were stretches of evaluative talk concerning the performance as a whole—tended to be lengthier and more discursive. A particular problem concerned how to deal with repeated references to the same aspect of the test-takers' speech within a single summary turn. Because it proved to be not always possible to identify whether repeated references to a category within a turn were repetitions of the same point or a new point, for the sake of consistency all references to one of the major aspects of the performance falling within a single unit were treated as a single unit. An ellipsis (…) was used to indicate the linking of two discontinuous pieces of text in the same ideas unit:

- Grammar: This speaker is rather simplistic in her grammar ... she doesn't seem to have a firm grasp of simple past tense. She doesn't seem to exhibit the use of any gerunds, so her grammar is at a fairly basic level.
- Pronunciation: Pronunciation was a little bit problematic in that I had to really focus to follow along at certain points here ... but there are some areas of weakness there in terms of pronunciation.

The definition of an ideas unit was revised, then, to deal with noncontinuous speech. It was redefined as "a single or several utterances, either continuous or separated by other talk but falling within the same turn, with a single aspect of the performance as the focus."

Before proceeding to the next level of analysis, intercoder agreement was calculated. In order to ascertain the relative stability of each of the categories, kappa was calculated for each one (see Table 2). This was done by collapsing the units into X and not-X for each of the categories

(e.g., *linguistic resources* and *not linguistic resources*) and calculating the level of agreement between the two coders. High levels of agreement, which ranged from 0.93 to 0.98, were achieved on all categories except *other.* As the *other* category had been used for all items that could not be clearly allocated to one of the main categories and thus was not a coherent category, it was expected that there would be a higher level of disagreement here than on the categories that could be relatively clearly described. It was also by far the smallest category.

Table 2

Intercoder Agreement by Major Category

	—Coder 1—						
Coder 2	Linguistic resources	Phonology	Fluency	Content	Global	Other	*kappa*
Linguistic resources	**1,535**	5	13	22[a]	0	2	0.94
Phonology	10	**847**	2	1	0	2	0.98
Fluency	9	2	**423**	6	0	1	0.95
Content	50[a]	1	3	**1,083**	7	2	0.94
Global	4	2	0	4	**150**	2	0.93
Other	0	2	2	0	1	**32**	0.79

[a] An area of disagreement that is discussed later in the Results section.

Once all disagreements at the upper level had been resolved through discussion, the coders turned to the second level of analysis, which involved finer levels of distinction within the major categories. Subcategories were determined by the two coders on the basis of iterative reviews and analyses of sections of the data. At this level of segmentation and coding, different aspects of the major categories were often syntactically interwoven in the speech of the judges so that it was not possible to physically segment the text into ideas units corresponding to the different subcategories, as it had been for the major categories. It was decided, therefore, that the most appropriate way to subcode the data would be to assign category codings to each unit for as many performance features as could reliably be identified by the two coders.

All ideas units belonging to each of the major categories were reviewed and coded in reference to the new subcategories, subdividing where appropriate to reflect reference to more than one feature within that major category. Again, a draft coding protocol consisting of descriptions of

the subcategories, including key terms and examples, was prepared. The coding procedure was then piloted by both coders, with attempts to apply the protocol to a subset of the data (again drawing selectively from all judges, tasks, and levels) followed by checking of agreement and amendment of subcategory descriptions and procedural instructions, as needed. Once a satisfactory level of consistency was achieved, the two coders each coded the entire data set.

Table 3 displays the levels of intercoder agreement achieved for the subcategories; intercoder agreement was calculated in the same way for each of the subcategories as it was for the major categories (i.e., through the construction of a two-by-two table with units recoded as X and not-X[3]). The level of agreement on each subcategory was above 0.90 for all the linguistic resources, phonology, and fluency subcategories, and above 0.80 for the content subcategories (see Table 3). Again, as expected the highest agreement was on phonological and fluency features, which were quite distinct and clearly identified in the speech of the judges.

Table 3

Intercoder Agreement by Subcategory

Linguistic resources	*kappa*
Grammar (sentence-level and below)	0.94
Vocabulary	0.95
Expression (undifferentiated grammar and vocabulary)	0.90
Textualization (above-sentence-level syntax/lexical markers)	0.94
Phonology	
Pronunciation	0.96
Intonation	0.97
Rhythm and stress	0.97
Fluency	
Hesitation	0.99
Repetition and repair	0.98
Speech rate	0.96
Fluency (nonspecific or global)	0.94
Content	
Task fulfillment	0.87
Amount	0.89
Ideas	0.80
Framing	0.88

Linguistic Resources

The largest group of comments pertained to test-takers' linguistic resources. At times judges distinguished *lexical resources*[4] or *vocabulary* from *grammatical knowledge*; at times they did not. In addition to commenting on what was essentially sentence-level grammar, judges also evaluated test-takers' linguistic resources with respect to the production of extended discourse, referring specifically to the use of connectives, discourse markers, and other cohesive devices.

Grammar. Judges made overall or global assessments of the test-takers' grammar, both in general terms (using terms like *basic*, *competent, limited*) and in terms of its adequacy for the demands of the task. More specifically, they focused on *accuracy* on the one hand and *sophistication* or *range* of structures used on the other. Accuracy was conceptualized in terms of the number or frequency of errors (Extract 1) and the ability to produce well-formed or complete sentences (Extract 2). The types of errors noted most frequently were tense (verb form and choice of tense in relation to functional requirements), subject-verb agreement, singular/plural marking, the use of articles, and less frequently, accurate or inaccurate formation of conditional/subjunctive structures, passive constructions, and relative clauses (Appendix C, Set 1, provides examples). More frequently than not, judges commented on the *impact* of errors on intelligibility (Extract 3). This would suggest that comprehensibility might be more salient to raters than accuracy *per se*. In terms of sophistication or range (Extract 4), features commented on most frequently included the use of complex or compound sentences, nominalizations, passive constructions, conditional and subjunctive verb forms, gerunds, and modals. Some of the structures referred to were task-specific. (Appendix C, Set 2, offers further examples.)

Extract 1: There are considerable grammatical inaccuracies.

Extract 2: Syntax, rather telegraphic at times.

Extract 3: And he is not always accurate in his language, a couple of places he's so inaccurate it's difficult to understand his point; well it's just barely discernable.

Extract 4: His sentence structures are complex and sophisticated.

Vocabulary. As was the case for grammar, judges made general assessments of test-takers' vocabulary skills and commented frequently on the adequacy of their vocabulary *for the particular task* (Extracts 5-6). They commented on the *accuracy*, *precision*, or *appropriateness* of specific lexical choices, as well as on the *sophistication* or *range* of test-takers' vocabulary in general (Extracts 7-8). As with grammar, judges were concerned also with the *impact* of vocabulary

choices on the intelligibility, clarity or precision of ideas (Extract 9). Aspects of vocabulary knowledge that received particular attention were correct choice of preposition, verb-preposition collocations, and the ability to produce correct word forms (i.e., adjective-adverb-noun transformations).

Extract 5: They certainly had the vocabulary range to be able to quite accurately describe the process, that experiment and its significance.

Extract 6: *They can get angry*: now the vocabulary is not sophisticated enough to express what the student is trying to say here. Word choice is not adequate.

Extract 7: You've got pretty basic vocabulary that is not being used very accurately or appropriately.

Extract 8: His vocabulary is quite rich.

Extract 9: … And then his generic *is important*. A better prepared student would've come out immediately with a position that would say something more than *important*; *important* doesn't really tell us anything, in that way.

There seemed to be conflicting views as to what constitutes evidence of lexical ability. At times judges looked for evidence of test-takers' ability to paraphrase or use alternatives to the lexis provided in the prompt or the input text (Extracts 10-11). In contrast were comments that appeared to value the ability to re-use the key terms provided in the input texts (Extracts 12-13).

Extract 10: And he's able to vary *summer vacation* with *holidays*, so he's not tied to the language of the task; he's got synonymous terms he can use for the same sense, ideas around the topic.

Extract 11: He nicely paraphrases here before he uses the term here—*they have the concept* of *numbers*—and that's his own paraphrase to explain the key item from the text.

Extract 12: Also in terms of vocabulary he had varied vocabulary and was able to pick up on the terms used in the lecture.

Extract 13: So again picking up the *cognitive ability* vocabulary.

A number of comments referred to stylistic choices in vocabulary. Close inspection of the comments on style, however, revealed conflicting perspectives among judges as to what was an appropriate style for spoken academic texts. Some comments indicated that formal, academic language was valued over colloquial word choice (Extracts 14-16), whereas others appeared to

value informal or colloquial language (Extracts 17-18).

Extract 14: He's got the vocabulary for the impersonal register.

Extract 15: *Kids* is not a good word choice, not very good register here.

Extract 16: *The students be around all year—around* is perhaps a little informal, but it's okay.

Extract 17: And then his use of *nerd* and *geek* is a further indication of his ability to manage informal register baggage. So overall, this is a student who's managing the informal register in relation to the topic. He's got the vocabulary which is appropriate to the topic.

Extract 18: I like that, there's such a sense of yeah, a [xxx] about that, *if we're stripped of our holidays*. Some people might say that that's an inappropriate use of it, but I actually think because of the authoritarian thing coming into it that he has used that term very idiomatically and very appropriately actually.

Expression. A relatively large number of comments made reference to linguistic resources in general, rather than distinguishing grammar from vocabulary. Judges used terms such as *language*, *expression*, or *resources*, again made global assessments, and referred to *sophistication* (Extract 19) or *control* (Extract 20). As for both grammar and vocabulary, they referred to the adequacy of the test-takers' language resources for the task (Extract 21), and to the impact of poor expression on comprehensibility (Extract 22).

Extract 19: Sophisticated language here as well.

Extract 20: So she'd had control generally throughout; she'd had control of what she wanted to say, and how she wanted to say it.

Extract 21: Overall, a lack of, I think, the language resources to construct an argument.

Extract 22: He has some problems of expression, and there were some times where I could not understand him.

Finally, as was the case with vocabulary, judges again referred to expression more generally in terms of its style. Again, while judges commented positively on features of academic speaking style (Extract 23) and on the use of colloquial or idiomatic language (Extracts 24-25) when they occurred, the use of lexical fillers and hedges that typify casual language use was not considered appropriate (Extract 26).

Extract 23: He seems to have a pretty good academic basis in his language and a good basis of academic discourse as well.

Extract 24: *That whole thing*—right—very colloquial and very relaxed and very natural speech.

Extract 25: *And as the years went on*, so very, very native-like colloquial expressions.

Extract 26: She uses some fairly informal colloquial language, for example *like* and *maybe*. In a performance of this kind or a task of this kind, I probably would not consider that acceptable.

Textualization. In addition to sentence-level grammar, judges commented on test-takers' ability to produce connected discourse. They commented on the use of what were variously termed *connectives, linkers,* or *cohesive devices* (Extract 27). They also occasional made references to cohesion problems with *missing referents* (Extract 28). Finally, they referred to the use of what were generally termed *discourse markers*—lexical items that explicitly staged the sections of the text (Extracts 29-30).

Extract 27: He is very good at linking these ideas, whatever these points are, you know with connectives—*because*, *if*, and *so on*.

Extract 28: Yeah, *maybe parents will welcome them*, that pronoun there has lost it's reference, it's too far back as a cohesive marker. *Them* is unidentifiable in this particular context. So, yeah perhaps an indication of some limitations to her ability to produce totally coherent discourse.

Extract 29: There were no linkers in there, no discourse markers such as "first," "second," to improvise an answer with reasons on the task.

Extract 30: She uses appropriate discourse markers to indicate the kind of information she's giving and to stage her information.

Table 4 summarizes the subcategories and aspects of performance to which judges attended within the linguistic resources category.

Table 4

Linguistic Resources: Summary of Raters' Orientations

Coded subcategories	Features noted by judges
Grammar	Adequacy for task Errors (types/frequency) Well-formed sentences Sophistication/range/complexity of structures Impact (comprehensibility)
Vocabulary	Adequacy for task Accuracy/precision/appropriateness of choices Sophistication/richness Stylistic choices Paraphrasing/re-use of prompt/input vocabulary Impact (comprehensibility)
Expression	Accuracy/control Sophistication/range Idiom Paraphrasing/re-use of input or prompt text Style Impact (comprehensibility)
Textualization	Connectives Cohesion Discourse markers

Phonology

Comments within the phonology category addressed *pronunciation* (the articulation of vowels and consonants) on the one hand, and prosodic marking in terms of *intonation* and *rhythm and stress* on the other. Pronunciation was by far the largest subcategory of phonology.

Pronunciation. Judges made overall or global assessments of test-takers' pronunciation (Extract 31). They also commented on its accuracy or nativeness (Extract 32) and, commonly, its impact on the intelligibility of test-takers' speech (Extracts 33-34). The most commonly identified types of errors included difficulties with "t"-"th" sounds; inability to produce final consonants in words, consonant clusters, and nasal-consonant clusters; lengthening or shortening of vowels; vowel shape and position; consonant substitutions; and insertion of epenthetic vowels (Appendix C, Set 3, provides illustrations).

Extract 31: Her pronunciation is reasonable.

Extract 32: Pronunciation is close to native-speaker standard.

Extract 33: There is however some difficulty with his pronunciation; on some occasions it's not possible to understand what he's saying.

Extract 34: Pronunciation is perfectly clear; there's no strain for the reader [*sic*].

Intonation. In comparison with pronunciation, relatively few comments specifically addressed prosodic features such as rhythm, stress, and intonation.[5] In terms of intonation, judges looked for native-like modulation (Extract 35). The impact of nonnative features on intelligibility was also noted (Extract 36). Judges also commented on test-takers' ability to break speech up into information units (Extract 37). This was regarded not only as natural, but also as aiding comprehension.

Extract 35: He's speaking very naturally … in a lot of his intonation.

Extract 36: And the intonation is pretty flat. There's not a lot of pitch change, so it's not easy to pick up or to really get clearly what he's saying. You can understand it, but there's a little bit of strain.

Extract 37: However his breaking up of information into units is really poor and so everything does just slide into the other, which makes it difficult for the listener. So therefore this student needs work on chunking things and intonation with information units.

Rhythm and stress. Judges commented on stress placement within words, particularly in relation to key task words (Extracts 38-39). Sentence stress (rhythm) was also commented on (Extracts 40-41), as was the placement of stress within utterances for emphasis (Extract 42-43).

Extract 38: Just a pronunciation comment—*experiment*—so faulty stress pattern there in his pronunciation of *experiment*.

Extract 39: But *numeracy* is the one word which is giving trouble. The stress is not correct.

Extract 40: He's also able to use sentence stress.

Extract 41: Because of the inaccurate rhythm—for example, *Fantz showed infant*—he's actually just reading out, I would say. So it's difficult to see where the groups, phrasing, to see any phrasing and therefore to understand what he's saying.

Extract 42: And the intonation is natural. There's effective use of emphasis.

Extract 43: Really good, very natural use of stress within the sentence. Stressing the fact that *it did help*: you know that's really quite native-like.

Table 5 summarizes the subcategories and aspects of performance to which judges attended within the *phonology c*ategory.

Table 5

Phonology: Summary of Raters' Orientations

Coded subcategories	Features noted by judges
Pronunciation	Accuracy/nativeness Impact (intelligibility) Specific errors Key words
Intonation	Naturalness/nativeness Impact (intelligibility) Information units
Rhythm and stress	Naturalness/nativeness Word stress Key words Sentence rhythm Emphatic stress

Fluency

The judges made a considerable number of references to the fluency of the test-takers' speech. At times they made overall evaluations of fluency (Extracts 44-45); at other times they referred to specific aspects of fluency—namely *hesitation* on the one hand (pauses and fillers) and *repair* on the other (repetition, rephrasing, and false starts).

Extract 44: From the point of view of language this speaker is close to a native speaker in fluency.

Extract 45: But certainly this is I'd say quite smooth flowing discourse.

Hesitation, pauses, and fillers. The largest number of comments within the fluency category were concerned with *hesitation* and *pauses* (Extracts 46-47). A number of comments also referred to the use of fillers (i.e., filled pauses; Extract 48). The majority of judges viewed fillers negatively, in terms of test-takers' inability to maintain their speech flow (Extracts 49-50). There were, however, occasional references to their naturalness (Extracts 51-52). Finally, the *impact* of

hesitation and pauses on intelligibility was a common concern (Extract 53).

Extract 46: It's very hesitant.

Extract 47: There's a lot of hesitations and pauses.

Extract 48: Quite a few *uh*s and *um*s.

Extract 49: She uses *um*, a discourse marker there correctly in a hesitation spot. But in her case, rather than an indication of, I think, successful management of informal English, it's another example of her inability to maintain flow. She's using empty space and *um* to fill gaps that she's not filling with content.

Extract 50: And the *um*s work against her authority in presenting this information.

Extract 51: He's using *er* a lot, possibly thinking on the run, but then it's not a disturbing aspect of his speech. It's something that native speakers do when they're thinking recall and putting it into their own words.

Extract 52: And it's quite interesting to hear her filler there—*I mean*—which is very native-like.

Extract 53: There was quite a lot of hesitation in there which again for the listener kind of causes a bit of strain because you're thinking "what's coming next?" and "how long is it gonna be till it comes?" sort of thing.

Judges made frequent inferences about the causes of hesitation. These included a concern with accuracy (planning the language), on the one hand (Extracts 54-55), and planning the content or organization of the information or ideas to be expressed, on the other (Extracts 56-57). The impression given by some comments (Extract 58) is that judges do not want to penalize test-takers for pauses arising from cognitive planning, because presumably, they see it as task-induced and therefore not truly representative of their proficiency.

Extract 54: And there are hesitations. And she seems to be torn between expressing the ideas that she wants to and forming correct sentences.

Extract 55: Again, long pauses which disrupt the flow of the discourse where I presume that the test-taker is searching for the correct word, again indicating a focus on accuracy at the expense of fluent discourse—fluent, flowing discourse.

Extract 56: Little bit hesitant, but her hesitancy seemed to arise from thinking about what she was going to say from the content rather than searching for the appropriate vocabulary in which to express it.

Extract 57: Long hesitation between the last phrase and her next response. So she's having difficulty with this recall and structuring.

Extract 58: She's very hesitant because I think she's trying to think about what to say rather than due to the fact that she may not be fluent per se. I think that her fluency's probably, it seems good, because at the times that she's not thinking she seems to be fairly fluent.

Repetition and repair. Another aspect of fluency that judges commented upon frequently was repair fluency, which refers to repetitions, rephrasing, and false starts. Judges observed the occurrence of repairs at the level of vocabulary (Extract 59-60), grammar (Extract 61), and content formulation (Extract 62). Attempts to repair speech were at times evaluated positively, being seen as an indication that test-takers were able to *monitor* their speech (Extract 63). However, generally the *disruption* to understanding was more salient (Extracts 64-65).

Extract 59: *Plan* and *proposal*—his repetition there, or restatement, little bit hard to know, could be an example of repair there. He's realized that *plan* isn't quite what relates to the task situation. He's restated as *proposal*, which is there in the task language, so he may be repairing in terms of sense or he may simply be using *proposal*, which is in the task item, as a crutch.

Extract 60: Yeah, here she's hesitating and stumbling over the expression "underground water:" *using a lot of underwater*.

Extract 61: He's initiated a repair, trying to—he's used the present tense at first and then repaired and used the future tense, so obviously he's aware of his performance, of grammar.

Extract 62: Yeah *they need to*, so we expect something to be forthcoming but then we seem to break down again. Oh, *they don't need to*, so we have a negative here. First *they need to*, then *they don't need to*.

Extract 63: So this student is quite—she's got a good ability to self-correct, so she's started with *be rewarded* but changed it to *get reward*. Quite competent use of language.

Extract 64: Again, those self-repetitions, self-corrections there—they're throughout this and it's very difficult actually to follow because of the preponderance of these.

Extract 65: Yeah, so there's focus on accuracy, self correction, self-repetition, which I find quite intrusive here.

Speech rate. Judges commented on the overall speech rate of test-takers, both slow and labored (Extract 66) and overly fast (Extract 67). Fast speech in particular *tended* to be attributed to nervousness (Extract 68). It did, however, often have an impact on intelligibility, as Extracts 67 and 68 show. Modulating the speech rate for emphasis (Extract 69) was also viewed favorably, although such comments were extremely rare.

Extract 66: He's a very slow speaker, very slow in trying to put his point across.

Extract 67: And sometimes the fact that he's too fast makes it extremely difficult to pick up what he's saying.

Extract 68: But she may be nervous; I dunno if she would speak like that all the time—possibly. But certainly it would have been a great response if she'd slow down a bit.

Extract 69: Speaking more slowly … Speaking more slowly seems to be for emphasis, which is appropriate.

Table 6 summarizes the subcategories and aspects of performance to which judges attended within the fluency category.

Table 6

Fluency: Summary of Raters' Orientations

Coded subcategories	Features noted by judges
Hesitation	Unfilled pauses/filled pauses/fillers Speech flow Naturalness Impact (intelligibility) Causes Linguistic (plan language/search for words) Cognitive (recall/plan/organize content)
Repetition and repair	Repetition/rephrasing/false starts Repair of grammar/vocabulary/content Self-monitoring Impact (intelligibility)
Speech rate	Slow/overly fast Causes (nervousness) Impact Emphatic slowdown

Content

As was expected given the academic nature of the tasks, the content of test-takers' responses was a major focus in both task types, independent and integrated. This category was second only to linguistic resources in terms of the number of comments. While the specific content expected varied from task to task, in this section we focus specifically on commonalities in the ways judges evaluated performance in terms of content across tasks. Subcategories identified here are termed task fulfillment, ideas, and framing. (Differences by task type are addressed later as part of research Question 2.)

Task fulfillment. Global assessments of the content of test-takers' responses were couched in terms of the completeness of responses (*task fulfillment*; Extract 70) and the extent to which responses were *on topic* or *addressed the question* (Extracts 71-72).

Extract 70: But they did fulfill the task to a certain extent.

Extract 71: So she has answered the question.

Extract 72: The task wasn't fulfilled, at least the task—I think maybe the task the student had in her mind was, but not what was actually asked here.

Ideas. In all tasks, judges commented on the amount of speech produced by test-takers, both in terms of number of ideas in general and in terms of sufficiency for the task (Extracts 73-74). Judges also evaluated the amount of speech relative to the time allowed (Extracts 75-76).

Extract 73: But overall he seems to have a very limited range of ideas to express on the topic as well.

Extract 74: Again, there's not a lot of detail given about the experiment, which I think is unfortunate.

Extract 75: So all in all she hasn't expressed much in the time allocated to her.

Extract 76: Well she finished with a very long pause, which means that she really didn't express that much in the time she was given.

Judges also discussed the content of the responses in terms of the ideas test-takers' constructed (independent tasks) or reproduced (integrated tasks). One way judges defined content was in terms of the *functional demands* of the task (Extracts 77-81). (As the exact functional requirements were task-specific, these are addressed in more detail during the discussion of Research Question 2.) They also commented on the quality of test-takers' ideas in terms of their *sophistication* (Extracts 82-83, independent tasks only) or *relevance* (Extract 84).

Extract 77: He was able to express an opinion.

Extract 78: She had a clear thesis, a clear opinion about the issue, and stated that, and then she listed three or four different values that those particular subjects had.

Extract 79: He clearly states the problem ... He details the cause, he states the cause very clearly, and then details the progression of events that led up to the current situation in this place.

Extract 80: So there's absolutely no information in there about the significance, the results of the experiment and the significance.

Extract 81: This speaker manages to identify the purpose of the experiment well. Talks about the reason for the experiment, is at the end of her presentation, identifies the causal link between time and perceptual discrimination. She also identifies the elements of the experiment, describing successfully, in my view, the three images.

Extract 82: But a reasonably sophisticated approach to the answer at least.

Extract 83: But in fact he's putting forward quite an interesting perspective—namely that, I think what he's saying is that, if courses are going to be cut, the cuts should be equal across all.

Extract 84: And also, that information, it's sort of peripheral really; it's not actually necessary to the task.

Judges commented on the *accuracy* of the information reported in responses to integrated tasks only (Extracts 85-86). As Extract 85 indicates, this could be interpreted as a listening problem and therefore not relevant to a test of speaking ability. The interface between listening and speaking is perhaps most obvious or salient at the level of the accuracy of information, but is also likely to be an issue in the selection and organization of information. (As this issue pertains to integrated tasks only, it is discussed in more detail under Research Question 2.)

Extract 85: That's a problem in terms of content. Because that actually wasn't what the experiment was showing, but I mean, that could be carrying over from listening comprehension.

Extract 86: You've also got some factual errors coming in again: errors, overgeneralizations that all the water comes through, groundwater and things like that.

Judges also referred to the logic or clarity of speakers' arguments, both at the level of individual points and at the level of overall organization or structuring of the text (Extracts 87-88). Overall organization of ideas was something that they frequently commented had an impact on the comprehensibility of the response (Extract 89).

Extract 87: Not too sure exactly what he's trying to say here in terms of following the logic of his argument here.

Extract 88: He's got his response quite logically ordered and structured.

Extract 89: Overall, really easy to follow, no strain because of the organization of the performance ... the processing that the student has done while listening to produce a shaped performance, basically.

Framing. There were some references to the generic structure of test-takers' responses in terms of an introduction and a conclusion (Extracts 90-92). Such comments were, however, relatively few in number, particularly in relation to the inclusion of a conclusion. Given the real-time nature of spoken performance and the limited planning and production time test-takers have compared to tests of written production, it may be that judges compensate for this by placing less value on structuring and framing when assessing spoken texts than they do when assessing written texts. (This issue is considered again during the analysis of speech samples in Study II.)

Extract 90: Well, this performance has a sense of introduction and conclusion, so it's quite well shaped.

Extract 91: And then she returns to the first point, the conclusion, so she has a conclusion—*it would be better*.

Extract 92: This speaker gives a good introduction.

Table 7 summarizes the subcategories and aspects of performance to which judges attended within the content category.

Table 7

Content: Summary of Raters' Orientations

Coded subcategories	Features noted by judges
Task fulfillment	Completeness of response On topic
Ideas	Amount (adequacy for task; proportion of speaking time used) Constructed/reproduced Functions addressed Sophistication (independent) Relevance of ideas (independent) Accuracy of ideas (integrated) Importance/relevance of ideas (integrated) Impact (intelligibility) Logic/clarity (individual ideas) Logic/clarity/coherence (text) Organization of ideas
Framing	Introduction Conclusion

Global Assessments

Some judges produced global or holistic assessments of test-takers' speaking proficiency (Extracts 93-94). It was not, however, always the case that the performance reflected a "flat profile" across all criteria (Extracts 95-96). Judges also evaluated performances in terms of the test purpose—namely to provide an indication of test-takers' readiness for academic study (Extracts 97-98).

Extract 93: This performance was excellent … And overall this was an excellent response that, and was, at a high level of speaking performance I thought.

Extract 94: I think less than satisfactory, but getting there.

Extract 95: So, as I said before, overall there's mixed levels for different aspects of the performance depending on the criteria that are going to be used.

Extract 96: So, approaching satisfactory in some respects, but not in others.

Extract 97: My feeling is that this is a student who would be struggling to cope with any type of tertiary study, any level.

Extract 98: And so my conclusion would be that she has the language skills, but she will need to practice them to come up to speed for participation in courses; otherwise

she will be left behind. She will also be perhaps tempted to plagiarize if she spends too much time in one area of a task, I think, whether writing or speaking. I say that because she seems to have the skill to do reasonably well with the language, having done that for the first part of the task, and therefore I'd suspect she will set herself that as her standard. This is not a well-balanced performance at all. It's a student with some language ability who would perhaps need assistance in terms of applying that language skill in a study context.

Table 8 summarizes the subcategories and aspects of performance to which judges attended within the global category.

Table 8

Global Assessments: Summary of Raters' Orientations

Coded subcategories	Features noted by judges
No subcategories	Holistic (flat profile) Analytic (marked profile) Readiness for study

Summary of Results for Research Question 1

In summary, the judges of test-takers' independent-task and integrated-task performances attended to four major conceptual categories when conducting their assessments: linguistic resources, phonology, fluency, and content. Within each category they considered a range of specific performance features. For example, within the linguistic resources category, they paid attention to grammar, vocabulary, expression (which included the use of academic or colloquial vocabulary), and textualization (which included text-structuring vocabulary such as discourse markers and connectives). Within the phonology category, they observed pronunciation, intonation, and rhythm and stress. For fluency, they noted hesitation, repetition and repair, and speech rate. And for content, they focused on task fulfillment (i.e., the completeness and relevance of the response or the extent to which the task was fulfilled); the amount, quality, and organization of ideas; and the framing of the response as a text.

Within these subcategories, judges tended to evaluate performance along a number of dimensions (shown in Tables 4-7). For example, in terms of linguistic resources, they referred to accuracy (the number or frequency of errors), sophistication or range (the complexity or difficulty

of the structures used), and impact (the effect of inaccurate grammar on the intelligibility or coherence of the response). Accuracy, sophistication, and impact are, of course, three separate dimensions on which any test-taker may show different levels of ability. For example, one test-taker may use a limited range of structures but have few errors, whereas another may use more complex structures and, as a consequence, make a larger number of errors.

In relation to phonology, judges focused on nativeness and intelligibility. Although intelligibility was referred to most frequently with respect to phonology, the impact of language problems on the intelligibility of test-takers' responses was a major concern with respect to all conceptual categories. For instance, in the linguistic resources category, judges made reference to poor grammar having an effect on the comprehensibility of test-takers' speech, and in terms of content, the selection or organization of the ideas or argument affected the clarity of the response. Judges also talked about the impact of disfluency (both hesitation and repair) on comprehensibility. The frequency with which judges referred to these various aspects of clarity (in addition to accuracy and sophistication) points to a potential tension in assessing spoken second-language skills—that is, the tension between nativeness, the standard by which learners are assessed against an idealized native-speaker norm, and comprehensibility, whereby the standard relates to the effort the listener has to exert to follow the speaker. This point is taken up further later in this discussion.

While there appeared to be general agreement as to what aspects of performance were valued, in some instances judges appeared to diverge in what they interpreted as an indication of proficiency. This was particularly the case in relation to the reuse of input text in the integrated tasks. Some judges, for example, appeared to value the re-use of input vocabulary in test-takers' responses, whereas others valued test-takers' ability to paraphrase or find alternatives for terms. Another source of disagreement among judges concerned the value they placed on disfluency associated with repair. While repair was generally judged negatively as interfering with comprehensibility, occasionally judges perceived repair as test-takers' ability to *monitor* and *correct* their own performances. In these latter instances, judges appeared to evaluate such disfluency positively. This finding echoes that of Brown (2000), who found that raters of IELTS interviews made positive comments about the use of self-correction, clarification strategies, and circumlocution by test candidates.

Although the largest category of comments pertained to linguistic resources, content, as

noted, was a major focus in the rating of both task types. We found that, although there was general agreement on the main functional and textual features sought in good responses, sufficient ambiguities and disagreements among judges as to what constitutes "key" information or important detail indicate that it may be necessary to provide raters, as part of rater training, with sample responses that indicate what content is expected for each task.

The fact that content is a major focus when judging performance on these types of tasks is a major finding of this study. Previous verbal-report studies involving raters of speaking-test performances—such as those by Meiron (1998), Brown (2000), and Pollitt and Murray (1996)—did not find content to be a major rating focus. However, these studies involved general or nonacademic tasks, such as interviews or monologic narratives, and in these contexts there seems to be a primary orientation to linguistic features, with the topic or content simply providing the vehicle for the demonstration of these linguistic skills.

McNamara (1996) makes a distinction between strong and weak performance tests on the basis of the criteria used in judging performances. In strong performance tests, performance "will be judged on real-world criteria, that is, the fulfillment of the task set" (p. 43), whereas on weak ones, the focus is the linguistic quality of the performance alone. The orientation to nonlinguistic aspects of performance here suggests that this test is at the stronger end of this continuum, and echoes the findings of a study of native-speaker readers on the IELTS reading test (Hamilton, Lopes, McNamara, & Sheridan, 1993), which demonstrates the role of general cognitive skills in performance on a test of academic reading. The heavy focus on the content of test-takers' responses in the context of EAP speaking tasks, even independent ones, indicates a clear orientation to the cognitive dimension of the tasks (i.e., the ability to produce well organized and meaningful content, as may be considered appropriate in a test with an academic focus). This reflects an orientation that resembles assessments of writing in an academic context (in which the raters tend to be explicitly directed to focus on the quality of the ideas or information produced; see Cumming et al., 2001). In this sense it differs from other assessments of speaking in which content tends to be less of a concern. It certainly indicates the need for content to be taken into account when developing scales for the assessment of cognitively complex speaking tasks.

Given the importance of text structure in academic contexts, the lack of a major conceptual category pertaining to textual or discourse features in the present analysis is worthy of comment. It is not that the judges did not talk about discoursal features of test-takers' performances—in fact,

there were many occasions in which they did. They did not, however, discuss discourse as a distinct category as they did other features, like fluency and pronunciation. They also did not make global assessments of discourse ability or refer to it otherwise as a conceptual assessment category. Discourse-relevant features of test-taker speech were generally attributed or attributable either to their linguistic resources (i.e., the ability to link information lexically and syntactically) or to the content of their responses (i.e., selection and ordering of information and framing of responses). Discourse ability, it seems, depends on having the linguistic resources necessary to create discourse (i.e., connectives, discourse markers, logical linkers), on knowing the generic or rhetorical structure required, and on having the cognitive ability to select or create and organize and sequence ideas or information.

Results: Research Question 2—Levels of Performance

Research Question 2 explores how judges differentiate levels of performance within conceptual categories. This analysis examined this question by extracting representative evaluative and descriptive statements from judges' verbal reports. In addition, following Pollitt and Murray (1996), we were also interested in whether all criteria were salient at all levels of performance. By identifying how field experts define increasing levels of EAP proficiency, we hoped to contribute information that will be of use in the validation of the existing draft scales for the new TOEFL exam, and more generally, to illustrate the usefulness of verbal report methodology for the development of scales or other scoring rubrics for the assessment of speaking tasks in other contexts.

As Table 9 shows, judges attended to all four major conceptual categories at all performance levels. For all proficiency levels, the majority of their comments fell in the linguistic resources category (30-40%), with content being the second largest category (25-30%). For these categories, the proportion of judges' comments increased slightly as proficiency level increased, while for fluency and phonology, the percentage of comments decreased with improved performance. Given that comments within the phonology and fluency categories tended to be negative rather than positive, the fact that the number of comments in these categories decreased at the higher levels is perhaps unsurprising. Marked differences in distribution across subcategories (e.g., grammar) and subcategory features (e.g., accuracy) are addressed through an impressionistic review of the data in the summary at the end of this section.

Table 9

Distribution of Comments Across Tasks

	Linguistic resources (%)	Content (%)	Phonology (%)	Fluency (%)	Global/ other (%)	Total (%)
Level 1	30	25	24	16	5	100
Level 2	35	25	22	14	4	100
Level 3	40	28	20	9	3	100
Level 4	40	27	20	8	5	100
Level 5	40	30	16	8	6	100

In order to answer Research Question 2, verbal report data were sorted by conceptual category, then by proficiency level, as defined by ETS ratings. We then examined the data, level by level, seeking to identify typical ways of describing different levels of proficiency. The findings are presented in the form of descriptive statements, one per level for each of the major conceptual categories. Each statement notes typical ways of evaluating performance; where possible, actual words used by judges are employed in these descriptions (again, indicated by italics).

Linguistic Resources

Judges described Level 1 test-takers as having *poor language control* or *poor grammar*, as evidenced by *many grammatical inaccuracies* and *inappropriate word choices*. Speech was described as *fragmented* with *many incomplete clauses* or *sentence fragments*. At times it consisted of *isolated words* and had *little or no tense-marking* as well as *basic tense and agreement problems*. Grammar and vocabulary were described as *not adequate for the task*. Test-takers tended to *repeat the task language* rather than paraphrase it and *struggled to formulate ideas*. Vocabulary was described as *extremely limited* or *restricted*; test-takers tended to use *high frequency vocabulary only,* and *basic word forms caused difficulty.* The impact was such that test-takers were *unable to get their meaning across, which caused great strain for the listener*.

At Level 2, grammar was described as *quite basic* or *limited*, and *errors were frequent*. Test-takers' responses were characterized by *some incomplete clauses or sentence fragments*; *basic tense problems*; and *little or no tense-marking*. Vocabulary was described as *not adequate for the task*, with test-takers having *difficulty finding the words*. *Cohesive devices and discourse markers* were *lacking*. Grammar problems were generally described as *interfering with*

understanding. Test-takers were *not always able to get the message across.*

At Level 3, test-takers' language was described as *fairly simple*. Expression *lacked sophistication*, and sentence structures were *simple.* Test-takers' responses were characterized by *some elliptical sentences* and *inadequacies* or *inaccuracies* in grammar, with *some basic grammar or tense problems*. Although test-takers' use of vocabulary included some *inadequacies* or *inaccuracies*, the responses employed a *reasonable level of vocabulary* and showed evidence *some ability* to *use vocabulary appropriate to the register*. they were also described as having *limited cohesion*. In general, speech was described as *not difficult to listen to*, with only *occasional strain for the listener*.

At Level 4, responses displayed *some* syntax and morphological problems or nonnative expressions, but test-takers were described as generally able to express themselves *quite well*. Grammatical or syntactic errors were described as *few* and *minor*, with only *occasional* tense problems. Test-takers used some *sophisticated* grammatical structures. Word order was *generally correct*. Test-takers employed a *good range* of vocabulary, with *some sophisticated* vocabulary or expression and *occasional awkwardness* in expression. *Good use* of connectives to link ideas and organizational (discourse) markers were also in evidence. Vocabulary was described as adequate for the task, and register was *generally appropriate*. At this level, test-takers were *able to* get their meaning across *clearly.*

At Level 5, test-takers were described as *articulate* or as having *no difficulty expressing themselves*. They had *strong control* of the language; errors were *rare,* and they used a *wide range of structures*. Grammar was described as *sophisticated*; test takers used *complex sentences* and showed *good use* of cohesion. Test-takers used a *wide range* of vocabulary that was *appropriate to the task*, *sophisticated*, *precise*, or *elegant*. They were *able to paraphrase* the task language, and their own language was described as *academically appropriate*, *native-like*, or *natural*. Errors *did not detract* from test-takers' meaning.

It was not possible to attribute specific errors to particular levels, partly because of the limited number of references to each type of error and partly because error types did not always appear at only one level. Basic subject-verb agreement problems were found even at Level 5, for instance. However, as expected, references to *fragmented* speech were common at Levels 1 and 2 (including references to the impact on meaning or intelligibility), less common at Level 3, and rare at Levels 4 and 5.

The question of *style* was addressed very little in lower-level performances; in other words, where test-takers' linguistic resources were very limited and very little speech was produced, style did not appear to be salient to judges. At Level 3, judges made a few comments on appropriate academic word choices and on inappropriate informal choices (such as *kids*). At Levels 4 and 5, they made many more comments, which dealt with both with academic and informal style, as noted earlier.

Phonology

At Level 1, pronunciation problems were generally described as *severe*, and errors were *frequent* and *intrusive*. Rhythm and intonation were often *extremely nonnative*, and speech was described as *monotone* or *staccato*. Stress patterns were frequently referred to as *nonnative* or *faulty.* The effect of these problems was that speech was often characterized as *unclear*, *difficult to understand*, or causing *a lot of strain* for the listener.

Level 2 test-takers were generally described as having *some pronunciation problems*. Nonnative or faulty stress patterns were *noticeable*, with *no or rare use of stress* to make meaning. In addition, misplaced word stress caused problems *at times*. Intonation tended to be *rather flat* or *unmodulated* with *not a lot* of pitch change. These problems were described as severe enough that decoding *required concentration* and speech was *difficult to understand* at times, causing *some strain* for the listener.

At Level 3, pronunciation was described as *reasonable* or *fair* with *occasional mispronunciations*. *Inappropriate or odd* stress patterns and intonation were evident *at times*. The impact was such that pronunciation was described as *nonnative but identifiable*, causing *occasional strain*. *Misplaced* stress or *inappropriate* intonation patterns caused *occasional difficulty* for the listener.

At Level 4, pronunciation was described as *not perfect*, but problems were *minor*. While test-takers used *occasional inappropriate* or *misplaced* stress, *occasional* use of emphatic stress was also noted. Although test-takers had *occasional problems* identifying words, articulation was described as quite clear, in that errors rarely interfered with intelligibility and didn't cause strain for the listener.

Level 5 test-takers were described as having native-speaker-like pronunciation with only the *occasional* mispronunciation. They had *strong* or *effective control* of intonation, stress, and rhythm, with stress and intonation also described as *natural* or *native-like*, and words described as

linked in a *native-like* way. Word stresses were deemed appropriate. In terms of impact, pronunciation problems caused *no strain* for the listener. Test-takers demonstrated *effective* use of intonation and rhythm to divide up meaning units and stage sections of text. They used stress *effectively* for emphasis and stressed key words *appropriately*.

Fluency

At Level 1, speech was described as *very hesitant*, with *frequent* or *long* pauses. Speech was *slow* with *constant* repetition or repair. The repetitions, repairs, and long pauses made speech *difficult to follow*. *Overly fast* or *slow* speech production *caused problems* for the listener.

At Level 2, speech was described as *rather hesitant*, with long pauses *at times*. The pace was generally *slow* or was described as having *erratic* or *jerky* fluency. Test-takers resorted to *a lot of* repetition or repair, with *constant* searching for words. Judges commented that there was a focus on accuracy to the detriment of fluency. The listener *had to concentrate* because of hesitancy. Disfluency was described as causing *quite a bit of strain* for the listener.

Level 3 speech was described as having reasonable fluency. It was a little hesitant with some pauses or occasional silence and *some* or *noticeable repair and repetition.* Disfluency caused *occasional strain* for the listener.

At Level 4, speech was described as *generally fluent*. Although test-takers displayed *occasional hesitation*, the flow of speech was *generally well-maintained*, and the *speech rate was good*. Hesitation was *not intrusive or disruptive,* self-corrections did *not intrude*, and problems caused *little strain* for the listener.

At Level 5 speech was described as *fluent*. The flow was *well-maintained* and test-takers responded with a *good speech rate*. Occasional hesitations, repetitions, or self-corrections were described as *natural* and *nonintrusive* and caused *no strain* for the listener.

Content

Increasing levels of content sophistication were described somewhat differently for each of the tasks and task types, making it substantially more difficult to define non-task-specific levels of proficiency for content than for the other categories. However, notwithstanding this complication, it was possible to extract general (i.e., non-task-specific) descriptors in relation to the following dimensions of proficiency:

- the extent to which test-takers addressed the task

- the amount of elaboration or detail evident in responses
- text structure
- logic/coherence/organization of ideas
- accuracy of information (tasks 3-5 only)
- relevance of ideas

At Level 1, the task was *misunderstood* or *addressed minimally*. *Little information* was presented; ideas were *few* and involved *little or no elaboration* or *detail*. The content of Level 1 responses was described as *not relating very closely* to the task, *not addressing all aspects* of the task, or as being *off the topic*. Responses contained *wrong* or *irrelevant information*, and test-takers were *unable to complete their responses*. Description or argument was described as *fragmentary*, *unsophisticated*, or *unsuccessfu*l. Content was *confusing*, and ideas were not clearly conveyed. Further, content consisted of *isolated points*, and logical connections were *not clear*. Information was *poorly* or *illogically organized.*

At Level 2, responses addressed a limited number of points. Ideas were described as *insufficiently developed*, *incomplete*, *simplistic*, or *inaccurate*. Major points were *omitted* or *distorted*, and test-takers used *insufficient detail*. The relevance of some points was *dubious* or *insufficiently explained*, and the logic of test-takers' arguments was *difficult to follow*. The organization of ideas was *not always logical* or *coherent*, and links between ideas were *not always clear*.

At Level 3, responses were described as *satisfactory* or *adequate*. While information was *generally well-organized*, some main or key points, details, or supporting information was missing. The argument was *not convincing* or *contained some logical flaws.* Some points were under-elaborated or inaccurate. The argument *lacked some clarity or coherence at times*.

At Level 4, responses were characterized by an *adequate amount of content*, and most information was relevant. The main points were *covered*, and supporting information or detail was described as *generally adequate*, although some may have been *missing, not fully elaborated*, *or inaccurate.* At this level, information was *generally well-organized*, the argument was generally *clear* or *coherent*, and links between ideas were *clear* or *logical*. The text was generally *well-structured*, with an introduction and/or conclusion, although some test-takers *may not have*

finished their responses within the allocated time.

At Level 5, responses were described as *accurate* or *thorough*. Main points were included, as were *ample supporting ideas* or *details*. Ideas were generally *sufficiently elaborated*, and information was clear and accurate. Responses were described as *well-organized* or *logically structured*. Ideas were *logically ordered* or the response *progressed logically* through all the steps of the content. The response or argument was *concise*, *clear*, *coherent* and *cohesive*. The text was *well-structured* with an introduction and conclusion.

Summary of Results for Research Question 2

Descriptors for the assessment of speaking performance tend to be relatively brief, general, and "superficial" (Cumming, Kantor, Powers, Santos, & Taylor, 2000). They have been criticized for the lack of empirical basis in terms of performance-related attributes and "relevance" to raters (Brown, 2000; Pollitt & Murray, 1996). As such, descriptors can be difficult to interpret and often require a level of interpretation that can result in potentially conflicting views of proficiency.

In general, we believe the results of this analysis show that verbal reports of the type derived here—which draw on experienced judges' unguided scoring orientations—do constitute an effective tool for the development of performance-level descriptors. By sorting verbal reports by conceptual category and proficiency level, it was possible to extract descriptive statements that reflect common ways of describing quality for the range of performance features referred to by judges. These statements, we argue, because of their basis in actual assessments by expert judges, will have a greater level of user-relevance than statements based merely on theoretical definitions of constructs or suppositions about which performance features might be relevant to judgments. (It was argued earlier that a comparison of descriptors derived from experts' verbal reports with those developed and piloted at ETS could provide empirical evidence of the user-validity of the operational analytic performance-level descriptors. This topic is taken up later in the discussion section, as Study II, which examines the extent to which trends identified in Study I are empirically supported in test-taker speech samples, is a further source of potential validation evidence.)

In undertaking this analysis, two features of the data were particularly noteworthy. First, the issue of holistic versus analytic ratings arose. When the data were sorted according to ETS-assigned ratings, we did *not* find it to be the case that all of the comments for any one category at any of the five rating levels reflected a consistent level of proficiency. However, given that the

ETS ratings consisted of a single holistic assessment, and that it is well known that learners often have "marked" profiles (i.e., they demonstrate a higher level of proficiency in some aspects of performance than others), all students who were evaluated holistically by ETS raters as Level 4, for example, were not necessarily expected to perform at Level 4 on all criteria.

For this reason, it was necessary to take the spread of their comments across levels into account when deriving "typical" descriptors. In particular, phonological sophistication did not always correlate well with the global assessment—as was the case, but to a lesser extent, for fluency. Also, by no means did speech at the highest level of proficiency contain no disfluency; however, this was relatively unsurprising, as judges frequently commented that disfluency was a feature of native as well as nonnative speech. It did not appear to be the case, then, that judges expected a completely fluent performance at the highest level.

In terms of the major categories, judges did indeed appear to attend to all aspects of proficiency at all levels performance—certainly in terms of the major categories. Within specific categories, however, certain aspects of performance were more or less salient at specific levels; these tended to relate to the production of sophisticated discourse, such as *style* (or register) and the *framing* of the response. This finding concurs with that of Pollitt and Murray (1996), whose raters focused less on discourse skills at the lower levels, and with that of Cumming et al. (2002) in the assessment of writing.

Finally, the evidence of the verbal reports suggests that performances assessed at Level 5 by ETS-trained raters did not always demonstrate the highest level of content quality. Given that judges at times commented on a disjunction between content and language,[6] when the two aspects of performance were not of the same (high) standard, it may well be that in the end, linguistic quality was privileged over content quality when assigning ratings.

Results: Research Question 3—Comparison of Tasks and Task Types

Research Question 3 investigates the extent to which the conceptual categories identified by EAP judges are task- or task-type-specific. Whereas Research Question 1 was concerned with commonalities across tasks and task types, Research Question 2 is concerned with differences. More specifically, this section addresses the question of whether separate scales might be warranted for independent and integrated EAP speaking tasks.

In general, as Table 10 shows, all judges attended to all conceptual categories across all tasks, and no marked differences in the distribution of comments by task type were observed. For

all tasks, the linguistic resources category (33-39%) was referred to most, closely followed by content (22-32%) and phonology (18-24%). Significantly, content was an important focus for independent tasks, as it was for integrated ones, indicating a task orientation rather than purely a linguistic orientation (Norris, 2001). For Tasks 3 and 4, the two listening-speaking tasks, the content of responses attracted a greater percentage of comments than the independent tasks and the reading-speaking task did. This finding may be a consequence of the fact that content was more problematic in those tasks in which test-takers' responses depended on their comprehension *and* their ability to remember the input. For both of these tasks, the percentage of comments devoted to purely production skills (pronunciation and fluency) were correspondingly lower.

Table 10

Distribution of Comments (Conceptual Category by Task)

Task	Task type	Linguistic resources (%)	Content (%)	Phonology (%)	Fluency (%)	Global/ other (%)	Total (%)
1	Independent	34	22	24	15	5	100
2	Independent	38	24	21	12	5	100
3	Integrated	39	31	18	9	3	100
4	Integrated	37	32	19	7	5	100
5	Integrated	36	24	21	13	6	100

An impressionistic analysis of the verbal reports was carried out in order to ascertain whether rating orientations specific to each of the task types—independent and integrated—were identifiable within each of the conceptual categories. As expected, the content category contained the most task-specific references, and these related both to the range and quality of ideas and the structure of responses. Also, some task-specific tendencies appeared within other conceptual categories, particularly linguistic resources (grammar and vocabulary). The major task- and task-type-specific differences found are described in the remainder of this section, starting with content; the implications for scale development are discussed later under Summary of Results for Research Question 3.

Content

The independent tasks required test-takers to present their own views on a particular issue: in one case (Task 1), the likely effect of an action, and in the other (Task 2), the value of a certain

educational offering. The analysis of verbal reports revealed that judges looked for a relatively restricted range of functions—namely, *opinion* supported with *reasons* and/or *examples*—as the overarching functional requirement of these tasks (Extracts 99-102). In the integrated tasks, the functional demands were more complex, but also task- (or more correctly, text-) specific.

Extract 99: And she's now expressing some supporting evidence for her opinion.

Extract 100: Now here he's attempting to—he's actually giving a reason for his opinion.

Extract 101: She had a clear thesis and lots and lots of support, explanations of values inherent in art and music and their uses.

Extract 102: This speaker gives specific examples of how art and music can benefit students.

Judges referred to the expected speech genre in integrated tasks in terms of *description, recount*, and *explanation* (Extracts 103-105). Within this general structure, they also referred to text-specific functional elements. For Task 3, for example, judges looked for ideas expressing what they referred to as *problem, cause, effect, solution*, *consequence,* or *reason* (Extract 106). For Tasks 4 and 5, judges looked for ideas expressing what they referred to as the *purpose/goal* of the experiment (or *the point of the experiment*), the *process*/*procedures* (or *the details*/*what happened*), the *results/findings,* and the *significance/conclusions* (or *what it showed*) (Extracts 107-108).

Extract 103: I could hear explanation, description of the problem, recounting in terms of explaining the stages in the problem in the valley.

Extract 104: But his description and recounting of the experiment, and his description of the elements of that, is good.

Extract 105: This is quite a good attempt to explain and describe the experiment.

Extract 106: She's presented the problem, the causes of the problem and the effect ... and presented solutions, and also has looked at the progressive worsening of the problem, and it's all stated very clearly in very well organized way ... I suppose in terms of information, she hasn't included—in talking about the cause of this problem—she hasn't included detail sequencing over time, how this developed over time.

Extract 107: There was certainly an attempt made to fulfill this task. There was a description of the details of the experiment, although there was a very low-

level accuracy, I felt, in terms of content of that experiment. And the result of the experiment—again there was an attempt made to convey the results and the significance, but a little bit of confusion there, I think, in the student's mind about exactly what those results showed.

Extract 108: The speaker explains, and describes, and recounts. There's a sense of the process or the procedure happening here, and the purpose is clear and the results of the experiment are also effectively summed up at the end.

In terms of the *structure* of responses, judges commented on the inclusion of an introduction and, less frequently, a conclusion, with respect to both task types. However, what this should consist of differed in the two types of tasks. Judges valued an opening statement of the issue in responses to independent tasks,, although an initial statement of opinion was also considered acceptable (Extract 109). A closing statement in which test-takers restated or summarized their opinions was also valued (Extract 110). While there was a general consensus that introductory and concluding statements were necessary in responses to integrated tasks, judges agreed less often on what these should contain. In Task 3 for example, some judges expected an introduction to the general issue of land subsidence (Extracts 111-112), whereas others were satisfied with an introduction to the specific instance described in the lecture (Extracts 113-114).

Extract 109: A clear expression of an opinion to start with.

Extract 110: And here he is drawing a conclusion—that if you were to sort of have an adverse affect from studying 12 months of the year, it would defeat the purpose of it.

Extract 111: And not too sure whether he's clear about what he's actually saying. He doesn't start by explaining anything about land subsidence, for instance.

Extract 112: This speaker launches straight away with the San Joaquin Valley as an illustration of land subsidence, but however he doesn't explain what land subsidence is.

Extract 113: She starts her introduction by saying *the problem in the San Joaquin Valley in California*, so she's going to identify the problem.

Extract 114: She's actually started her explanation quite well in that she's been able to fairly clearly and directly explain the nature of the problem and when it occurred.

In terms of the quality of the ideas expressed in the performances, in the independent tasks judges referred to the *sophistication* of the ideas (Extracts 115-116) and their *relevance* to the task (Extracts 117-118). As integrated tasks provided content though the input text, *sophistication* of ideas was not an issue as it was for the independent tasks. In integrated tasks judges looked for the inclusion of specific *key information* from the input texts, which they described in terms of *main ideas*, *key points* or *concepts*, and so on (Extracts 119-21).

Extract 115: *Spend their time playing jokes on other people*: Again, it's not the most sophisticated example to give. Perhaps it would've been better to describe vandalism or something like that.

Extract 116: Well, I guess the point that she's making here is the point about students not being happy. This seems a little simplistic.

Extract 117: He's using appropriate examples.

Extract 118: But it's really difficult to see what point he's making and how—particularly the second half of the performance or the examples he's giving—how that actually supports his approval of this measure. And in fact, it seems to be a contradiction when he claims that *we will have more time to spend with friends and do other things*, when in fact students will be studying longer.

Extract 119: So just to sum up, he's got a reasonably good view of the lecture. He's been able to summarize the key points in it with some accuracy and to recount those points.

Extract 120: She didn't really pick up the main point of the dialogue, which was really about the relationship between language ability and numbers, if I'm not mistaken, and differentiation of numbers. She seemed to think it was just all about a comparison between language learning abilities of monkeys and humans.

Extract 121: So she's picked up on the main point here, that the monkeys have numeracy ability. That was the point of the experiment.

In responses to integrated tasks, judges also commented on the *appropriateness* or *relevance* of particular items of information selected from the input text (Extract 122) and the *accuracy* of the information provided by test-takers (Extract 123). The *organization* of information was also more salient in responses to integrated tasks (Extracts 124-125).

Extract 122: The speaker manages to address the three sections of the task and to include some relevant detail.

Extract 123: Most of the information is correct, although there was a major inaccuracy in terms of 1970s and 1990s getting mixed up there.

Extract 124: But things are organized rather poorly. He's not actually organizing the information very well for the listener—so that if I did not know the lecture and know the information that he was trying to tell me, I would probably find that really difficult to follow.

Extract 125: He's really recounting what he's heard, rather than organizing the information to meet the requirements of the task.

Some discrepancies were noted in the value assigned by different judges to certain aspects of the content of test-takers' performances. In one of the independent tasks, for example, while the task asked specifically for the effects on the community, some test-takers addressed the effects on individual students. Judges appeared to differ in the extent to which they found the more personal approach acceptable (Extracts 126-127). In Task 4, judges' expectations about what comprised an accurate response were also not always the same: Some commented on the test-taker's ability to refer to the relationship between language ability and numeracy, while others were more interested in whether the test-taker had included the point that monkeys have an understanding of the concept of numbers (Extracts 128-129).

Extract 126: He does not address the affect on the community of holding school for 12 months of the year. He addresses only the effect on the individual student.

Extract 127: This student seemed to interpret the prompt in a very personal way, so the response was one of empathy almost, was not that level of abstractional normalization that you might want, in terms of the broader impact on the community. It was very much a "how this would affect a student, and I'm a student, so I can kind of empathize with that" sort of response ... but I think it was a very satisfactory response from a personal take on the prompt, rather than the sort of more abstracted response.

Extract 128: She could've slightly expanded on, at the end, how this proof of numeracy in rhesus monkeys, how that connects then to assumptions about language ability

and numeracy as described in the lecture, and therefore the overall conclusion based on that.

Extract 129: Yes, there seems to have been a lot of time spent here in this latter part of the student's description on the conclusion about the connection between language ability and ability to count, rather than on the actual conclusions that were drawn from this experiment, which were about the ability of monkeys or their understanding of the concept of numbers.

Linguistic Resources

While judges did not appear to make reference to any task-type-specific grammar or vocabulary, they did refer to the use of task- or text-specific structures and vocabulary. These were, however, few in total. They included (Extracts 130-133) the appropriateness of particular tense choices in relation to the functional demands of the tasks (all tasks), the use of the *conditional* (Task 1 especially, but also Tasks 2 and 4), and the ability to use *gerund* constructions (Tasks 1, 2, and 3). The use of the *passive,* which could be expected in responses to Tasks 4 and 5 in particular (integrated tasks concerned with reporting experiment details), was commented on across all tasks with the exception of Task 2. Other task-specific features included the use of the infinitive/present participle in constructions such as *to stop doing* and *to start doing* in Task 3 for the description of steps in a sequence of actions.

Extract 130: There's an attempt here at using a conditional.

Extract 131: *Will be raised* too: future passive correctly used there.

Extract 132: *The monkey can be reward*: now okay, got the idea of the passive, but the past participle is wrong and the verb choice is wrong—should've been "was rewarded" or "would be rewarded."

Extract 133: And she should've used the past tense. She doesn't seem to be able to express this in the passive. … I would [be] looking, at this level, for perhaps the use of the passive in this type of task.

While judges commented on vocabulary sophistication/breadth and appropriateness of word choice across all tasks, in the independent tasks they commented on the relationship between test-takers' lexical resources and their ability to express their own ideas. Such comments were most common for lower-proficiency performances, about which judges speculated that limitations in vocabulary prevented test-takers from expressing more complex or sophisticated ideas. In

contrast, in the integrated tasks judges focused on test-takers' ability to accurately re-use key terms provided in the input text (Extracts 134-135), although they at times also valued test-takers' ability to paraphrase the provided terminology (Extracts 136-137), particularly in the reading-speaking task (Task 5).

Extract 134: Yeah, I think she's used *inborned* or *inborn*, and I think she means "innate," so there's a problem with her use of words there.

Extract 135: Okay he starts with *the experience is about*, instead of "the experiment," so that's [a] vocabulary issue here.

Extract 136: Now here we have a very good, competent paraphrasing—*overpumping of underground water*. This composite noun of *overpumping*, very sophisticated.

Extract 137: So he's using key vocabulary from the reading in his own way, paraphrasing a little, which is good.

In terms of expression in general, in integrated reading-speaking task judges were also concerned with test-takers' ability to paraphrase input text rather than reproduce it exactly (Extracts 138-139). This was not an issue in the independent tasks, except when judges commented that test-takers reproduced expressions from the prompt,[7] which was generally interpreted as an indication of limited ability (i.e. the inability to move beyond the language of the prompt; Extracts 140-141). Such comments were very frequent at the lower levels, but less so at the upper ones.

Extract 138: This is lifted straight out of the text, the last line of the first paragraph. So she's using the text as a support to get through, because she can't restate in her own terms.

Extract 139: And using the exact words from the reading.

Extract 140: Most of her initial response, two-thirds of it at least, is simply quotation from either the speaking-task card or the reading passage itself.

Extract 141: She's used the task language rather than being able to paraphrase.

Phonology

The only noticeable phonological difference between task types was that judges commented on difficulties in pronouncing key terms in responses to integrated tasks, although such comments were few in number (Extracts 142 and 143).

Extract 142: She stumbles a little over the foreign pronunciation of San Joaquin Valley, which possibly affects her confidence a little.

Extract 143: The key word subsidence—mispronounced of course.

The terms "Joaquin" (Levels 3 and 4), "subsidence" (Levels 1-4), and "California" (Levels 1 and 2) in Task 3 appeared to be the most problematic. Whether this had an impact on the overall quality of pronunciation in this task will be examined in Study II, the speech samples analysis. It is also worth noting that the difficulties, at least with "Joaquin," can be attributed to the fact that this is an unfamiliar place name to many test-takers, which may also point to a bias toward certain language backgrounds.

Fluency

Although there were no noticeable differences between task types in terms of the aspects of fluency on which judges focused, differences in the number of references to the *causes* of disfluency were observed. In responses to integrated tasks, judges made more references to *cognitive planning* (planning the content of the response) or *recall of input* (Extract 144-145) as possible causes of disfluency, which might indicate that a higher level of task-induced disfluency is possible when test-takers respond to integrated tasks than when they respond to independent tasks. In other words, test-takers may be less fluent when completing the more cognitively complex integrated tasks. This hypothesis is examined further in Study II.

Extract 144: Well, the fact that there's this big gap where she's trying to get her thoughts together and frame or formulate an explanation probably indicates that she has not really understood the explanations provided in the lecture.

Extract 145: Little bit hesitant, but her hesitancy seemed to arise from thinking about what she was going to say from the content, rather than searching for the appropriate vocabulary in which to express it.

Summary of Results for Research Question 3

An examination of Research Question 3 showed that judges attend differently to the content and linguistic resources categories depending on task-type, which would seem to imply the need for distinct criteria for the two types of tasks. For content, at the most general level the criteria for independent tasks would need to refer to the quality of ideas in terms of their *relevance* and/or *sophistication*, whereas the criteria for integrated tasks would need to refer to *accuracy* and

the inclusion of *key information* and supporting *detail*. The *organization* of ideas appears also to be more salient in integrated tasks. For linguistic resources, in addition to *accuracy* and *complexity*, judges commented on the *re-use* or *paraphrasing* of input text. It would, therefore, appear to be necessary to refer to this ability in the scales themselves. This feature can be expected to have a counterpart in the assessment of academic writing using integrated tasks.

While differences pertaining to the abovementioned categories would appear to call for separate scales, as they reflect criteria that are substantially different in focus, for the other two categories the differences appear to be more a matter of degree. For pronunciation, for example, the only difference was that judges commented on specific key terms; however, the essential focus on *nativeness* or *intelligibility* remained the same. Similarly for fluency, the only difference was that judges talked more about cognitive planning as a cause of disfluency in the integrated tasks. However, such inferences about causes do not, in themselves, have any implications for scale development.

Judges also talked about performance in relation to the use of task- (or text-) specific items (which included particular functions, structures, or vocabulary), but also, for the integrated tasks in particular, the particular information expected in responses (key ideas and supporting detail). Given that the evidence here suggests that raters may not always agree as to what information is relevant or appropriate, it would be helpful to provide them with a description of the expected schematic structure (illustrated later in Figures 1 and 2) and informational content of responses for each task at each score level, along with sample responses, possible annotated. Similarly, for the independent tasks it may be helpful to provide raters with annotated sample performances that illustrate what is meant by the terms *sophisticated* and *relevant* in order to ensure equivalence of expectations.

Study II: Speech Samples Study

Introduction

The aim of the speech sample analysis was to explore the extent to which a correspondence exists between judges' perceptions of candidates' performances across proficiency levels and task types (Study I) and measurable features of the candidate discourse itself. It was important, therefore, that the analyses focus as closely as possible on performance features identified in the rater cognition study.

In order to identify methods of analysis appropriate to the conceptual categories that

emerged from the rater data, relevant literature on discourse analysis and interlanguage analysis in studies of second language acquisition was reviewed and linguistic experts were consulted. A phonetician was especially informative with respect to production features. A number of measures were identified, which are described in the next section.

All 200 speech samples were transcribed using transcription guidelines described in a study by Ortega, Rabie, Iwashita, and Norris (1998). The segmented and coded speech samples were entered into a database for use with the Computerized Language Analysis (CLAN) program developed for the CHILDES project (MacWhinney, 2000). The CLAN program allows a large number of automatic analyses to be performed on the data, including frequency counts, word searches, co-occurrence analyses, and the calculation of type-token ratios.

Methodology

This section introduces each of the measurable features analyzed in the speech samples; they are grouped according to the conceptual categories used in the rater study. In subsequent sections, results are reported for each feature—first in terms of differences that pertain to proficiency level and then in terms of task type. The information on measurable features provided in this section is necessarily quite detailed, as it is important to characterize exactly the features identified and explain various technical procedural decisions.

Linguistic Resources

The analysis of linguistic resources focused on four features: grammatical accuracy, grammatical complexity, textualization (the use of logical connectives), and vocabulary. Each of these is discussed in turn the paragraphs that follow.

Grammatical accuracy. Empirical studies in both language testing and second-language acquisition have reported measures of grammatical accuracy of learner speech in terms of either:

- global accuracy (i.e., identifying any and all types of error; e.g., Foster & Skehan, 1996; Skehan & Foster, 1999)
- specific types of errors (e.g., Ortega, 1999; Robinson, 1995; Wigglesworth, 1997)

The global-accuracy approach has the advantage of being potentially the most comprehensive in that all errors are considered. However, it is also the hardest in which to establish consistency of coding. Perhaps symptomatically of this problem, in the studies of global accuracy in Skehan and

Foster (1999) and Foster and Skehan (1996), no intercoder reliability measures are reported. An earlier study for the TOEFL project (Iwashita et al., 2001; McNamara, Elder, & Iwashita, 1999) also found that coders tended not to agree on what they considered to be errors, or on whether certain errors should be classified as grammatical or lexical errors. Specific types of errors, on the other hand, do not involve the same problem of reliability, but are narrower and less inclusive of all the potential features of accuracy. Given these uncertainties, a decision was made to measure grammatical accuracy through both methods: global accuracy and accuracy of use of specific grammatical features.

The specific features examined were use of verb tense, third-person singular, plural markers, prepositions, and articles—all features on which judges commented relatively frequently in Study I. Table 11 details how each of these grammatical features was defined and measured. Global accuracy was examined by calculating error-free T-units as a percentage of the total number of T-units. A T-unit was defined as an independent clause and all its dependent clauses (Hunt, 1970), while error-free T-units were considered to be T-units that were free from any grammatical errors, including both the specific errors defined above as well as any others (e.g., errors involving word order, omission of pronouns, etc.).

Errors were identified using the "target-like-use" analysis developed by Pica (1983) rather than a "supplied in obligatory context" analysis. The difference here is that the target-like-use analysis includes learner errors produced in both nonobligatory contexts and obligatory contexts. In counting errors in (and correct use of) the features listed above, the transcribed speech was first pruned by excluding features of repair. This meant that learners were considered to have shown evidence of correct use of the target-like feature when it was demonstrated in their *repaired* utterances.

Grammatical complexity. When judges referred to the complexity of learner speech, they were concerned with two general aspects. The first of these was the complexity of utterances at the level of clause relations—that is, the use of conjunctions and, in particular, the presence of subordination. We call this aspect "sentence complexity." The second was the perceived difficulty (or sophistication) of the structures used; for example, constructions involving passive forms, modal verbs, and comparatives were seen as "difficult" or "sophisticated." We call this aspect "sophistication." The analysis thus takes both of these aspects of grammatical complexity into account.

Table 11

Descriptions of Specific Grammatical Errors

Error type	Description
Tense-marking errors[a]	• Omission of the past tense morpheme (i.e., -ed) • Use of the base form[b] of an irregular verb/copula/auxiliary instead of a past tense verb/copula/auxiliary (e.g., "sink" for "sank," "is" for "was," "do" for "did") • Use of the base form of a verb/copula/auxiliary where future tense is expected to be used (i.e., omission of auxiliary "will") • Use of the base form of a verb instead of the passive form (e.g., "pump" for "was pumped") • Use of the base form of a verb instead of the progressive form (e.g., "increase" for "increasing") • Use of the base form of a verb instead of a gerund or participle (e.g., "stop pump" for "stop pumping," "reduce pumping by import water" for "reduce pumping by importing water")
Third-person-singular verbs/copula	• Omission of the third-person-singular morpheme (i.e., -s, -es) • Use of incorrect copula (e.g., "is" instead of "are"), irregular third-person-singular verbs (e.g. "have" instead of "has")
Plural nouns	• Omission of the plural (i.e., -s) • Use of a singular noun where a plural noun is required (e.g., "child" for "children")
Article use	• Omission of indefinite and definite articles • Incorrect use of articles (i.e., use of the definite article for the indefinite article and vice versa)
Prepositions	• Use of an incorrect preposition • Omission of a preposition • Use of preposition in nonobligatory contexts

[a] Included here are errors on which judges typically commented using the term *tense* (including passive/active voice and aspect).

[b] The term "base form" was used here instead of "present tense/infinitive" to refer to the uninflected form of the verb/copula/auxiliary because it is the most basic (i.e., unmarked) form and its use cannot be assumed to reflect a conscious tense choice by learners.

Sentence complexity was measured four ways:

1. by calculating the number of clauses per T-unit (the T-unit-complexity ratio)
2. by determining the ratio of dependent clauses to the total number of clauses (the dependent-clause ratio)
3. by counting the number of verb phrases per T-unit (the verb-phrase ratio)
4. by assessing the mean length of each utterance

The first three of these measures were identified in a review of second-language writing studies by Wolfe-Quintero, Inagaki, and Kim (1998) as the best measures to capture grammatical complexity, and have also been used in studies involving the analysis of learner speech in both pedagogic and testing contexts (e.g., Iwashita et al., 2001; Skehan & Foster, 1999). The T-unit-complexity ratio measures how grammatically complex learner speech is by assuming that the more clauses there are per T-unit, the more complex the speech is. The dependent-clause ratio and the verb-phrase ratio both examine the degree of embedding in a text—the former by counting the number of dependent clauses as a percentage of the total number of clauses, and the latter by counting the number of verb phrases per T-unit. The mean length of utterance was initially developed to investigate syntactic development in child language-acquisition studies (e.g., Brown, 1973), but some second-language studies (e.g., Ortega, 1998) found it to be a good measure for separating higher-level learners from lower-level learners. Mean length of utterance was measured by calculating the number of morphemes per utterance, with an utterance including both T-units and fragments.

Segmentation into T-units and clauses was undertaken according to guidelines developed by Ortega, Iwashita, Rabie, and Norris (1998) for an ongoing study of learner discourse. According to these guidelines, a T-unit is defined as an independent clause and all of its dependent clauses (Hunt, 1970); a clause is a unified predicate containing a finite verb, a predicate adjective, or a nontarget-like predication in which the verb or part of the verb phrase is missing (Berman & Slobin, 1994); and a dependent clause is a unified predicate (i.e., containing a finite verb, a predicate adjective, or a nontarget-like predication in which the verb or part of the verb phrase is missing) embedded in or dependent on a main matrix clause. With this approach, clauses can be

target-like or nontarget-like and complete or incomplete, but semantically they must constitute a single predication concerning a single activity, event, or state. Thus, coordinated clauses each count as different T-units and subordinated clauses together with the main clause count as a single T-unit. Verb phrases can be adjectival, adverbial, and nominal under these guidelines, depending on their function. While all clauses are in fact verb phrases, including those nontarget-like clauses that are missing the verb or part of it, together with verbless clauses (such as elliptical and gapped clauses), in the present study only nonfinite verbs (i.e., bare infinitives, to-infinitives, gerunds, and gerundives) were coded as verb phrases.

In measuring sentence complexity, it was determined that accuracy and complexity should be treated as distinct dimensions. Thus, all T-units, clauses, and verb phrases—whether or not they contained errors—were included in the complexity analysis. (Given that most T-units and clauses in the analyses contained some kind of error, if all of these had been excluded, very few would have remained for analysis.) Sophistication is measured in terms of simple frequency counts per 100 words for modals, comparative forms, and passive forms.

Textualization. The examination of test-takers' use of logical connectives was based on the work of Martin (1992), who categorized conjunctions at three different levels. Table 12 summarizes these categories, as described by Paltridge (2000), and provides examples of each type. As the table shows, at the first level there are four types: additive, comparative, temporal, and consequential. Additive conjunctions draw on the notion of addition in both a positive and contrastive sense; comparative conjunctions present comparison in both a positive and negative sense; temporal conjunctions are used to clarify the sequence of events expressed in a sentence; and consequential conjunctions refer to the consequential order of events expressed in a sentence. At the second level, conjunctions are grouped into two types, internal or external, in terms of interclausal/intersentential cohesive functions: Internal conjunctions relate to the staging of a text, while external conjunctions connect clauses. The third level of analysis is conducted in terms of the relationship, hypotactic or paratactic, between two clauses linked by a conjunction: Hypotactic relationships refer to clauses that are in an uneven or subordinating relationship with each other, while paratactic relationships involve clauses which are in a even or coordinating relationship with each other. Any single conjunction can be categorized multiple times, possibly at each of the three levels.)

Table 12

Categories of Conjunction

Category	Examples
Level 1	
Additive	And, or, moreover, in addition, alternatively
Comparative	Whereas, but, on the other hand, likewise, equally
Temporal	While, when, after, then, meanwhile, finally
Consequential	So that, because, thus, since, if, therefore
Level 2	
Internal	That is, for example, in fact, on the other hand, in conclusion
External	And, because, after, while, when, since, before
Level 3	
Paratactic	I tidied up my desk. It needed it.
Hypotactic	I tidied up my desk because I couldn't find the agenda.

Note. See Paltridge (2000), p. 137.

In the speech sample analysis, a total of nine different types of connectives were analyzed, as follows:

Connective	Level 1	Level 2	Level 3
and	additive	external	paratactic
but	comparative	external	paratactic
then	temporal	external	paratactic
so	consequential	external	paratactic
and	additive	internal	paratactic
but	comparative	internal	paratactic
when	temporal	external	paratactic
because	consequential	external	paratactic
not only … but also	comparative	internal	hypotactic

The connectives "and" and "but" were each classified into two different types: "And" could be additive/external/paratactic and additive/internal/paratactic, and "but" could be comparative/external/paratactic and comparative/internal/paratactic. "And" was considered external if it linked two parts of the same structural unit (e.g., the same process or outcome), but internal if it linked two different units (e.g., a process to an outcome; see Appendix D, Table D1, examples 2, 3, and 7). However, the distinction was not always clear-cut. When "and" connected two independent clauses, it was regarded as "external" even if the clauses were not in the same structural unit (see Appendix D, Table D2, example 17). Due to the complexity of the analysis, only 30 performances from Task 2 (independent) and 40 performances from Task 3 (integrated) were analyzed. (Appendix D provides analyses for two speech samples.)

Vocabulary. Vocabulary knowledge was examined using the Web-based program, VocabProfile (Cobb, 2002), which measures the proportions of low and high frequency vocabulary used by speakers, both native and nonnative. The program is based on the Vocabulary Profile (Laufer & Nation, 1995) and performs lexical text analysis using the Academic Word List (Coxhead, 2000). In addition to calculating measures for word-token, word-type,[8] and the type-token ratio, VocabProfile calculates the percentage of words in each of four categories: the most frequent 1,000 words of English, the second most frequent 1,000 words of English, words found in the Academic Word List, and any remaining words.

In order to undertake the analysis, the transcribed speech was pruned to exclude features of repair and imported into VocabProfile.[9] Frequency counts were then developed for each of the seven measures listed above. The word-token measure was used because it was assumed that, for weaker candidates or performances on tasks that were more cognitively demanding, not all of the allowed time would be taken up with speech, and even if it was, it was likely to be slower and thus yield fewer tokens. The word-type measure was chosen as a gauge of the range of vocabulary used; it was hypothesized that more proficient speakers and speakers addressing more cognitively complex tasks would use a wider range of word-types. In order to enable comparisons across tasks with different times allowed for completion, instances of word-tokens and word-types were counted per 60 seconds of speech; this was not necessary (and was not done) for comparisons across different proficiency levels. It should be noted that since a limited length of speaking time is allowed for each task, candidates who can speak fast produce more word-tokens than candidates who speak slowly. For that reason, the number of word-tokens could be affected by the speed of

candidate speech. The type-token ratio is a measure of the semantic density of speech, which may vary according to task or proficiency level; increases in tokens across levels may not be matched proportionately by increases in types, so speech may be relatively more or less dense at different levels or across different tasks.

Phonology

Prior to the analysis, advice was sought from phoneticians regarding the measurement of production features. The analysis focused on pronunciation, intonation, and rhythm, all of which judges commented on in the rater cognition study. Each of these is discussed in turn in the paragraphs that follow.

Pronunciation. The analysis of pronunciation features was conducted at word-level and subword-level. In the word-level analysis, coders first categorized words as meaningful or not meaningful. The pronunciation of meaningful words was then classified as target-like, marginally nontarget-like, or clearly nontarget-like in terms of morphophonemic ending and stress. The nonmeaningful words were categorized into three types: filled pauses (such as "um" and "er"), incomprehensible words, and words not completed.

In the subword-level analysis, syllables were again assessed as to whether they were target-like, marginally nontarget-like, or clearly nontarget-like. In addition, those syllables which were not identified as being in any of the three categories were assessed as epenthetic syllables, which were further judged as clear or marginal. The character of the nontarget-like syllables was assessed in terms of four features: aspiration (whispered/swallowed), pronunciation of <th> (e.g., this, that, there), consonant quality, and vowel quality.

Intonation. The assessment of intonation was conducted in terms of the number of completed intonation units.[10] The analysis considered whether test-takers produced completed units, cut off or incomplete units, or isolated words. Performances were first categorized as displaying many or few intonation units. The many category was then further broken down into performances showing English-like intonation, nearly English-like intonation, and non-English-like intonation. Criteria for allocation into these subcategories included: Do they follow general patterns, such as rising pitch to indicate continuation and falling pitch to end a thematic section? Do they place pitch accent on focused words and phrases in the sentence? Do they pronounce their English using the intonation patterns of another language?

Rhythm. Most varieties of English have a rhythm based on word stress, so that stressed

syllables come at a regular rate (i.e., they are stress-timed). In contrast, many other languages and even some varieties of English (e.g., Indian, Singaporean) are syllable-timed: Generally, each syllable comes at a regular speed. Syllable-timed speech is known to be particularly problematic for speakers of stress-timed languages, and vice versa. The categories used for the analysis of rhythm include stress-timed, syllable-timed, variable (for speakers who wavered between the two), and unclear (when judges could not really tell, which tended to happen when speech sections were not long enough).

The analysis was undertaken using the speech-data labeling software, Xwaves (Rommark, 1995). Xwaves was used because it allows a number features to be transcribed against a recording, tagging them to particular time-points. The frequency of each type of tag can then be calculated. Also, the Xwaves program includes a labeling module, which can be set up as one likes. Comments or labels for words and any other features of interest can be entered onto the screen in the form of ASCII text and aligned with a particular point in the speech file (labels for words, segments, and the like are conventionally aligned with the right edge of the unit). Because of the time-consuming nature of the analyses they were carried out on a portion of the data only: 30 seconds from each performance on one independent task (Task 2) and one integrated task (Task 3).

Fluency

The following features were identified as suitable measures of fluency: filled pauses (ums and ers), unfilled pauses, repair, total pause time (as a percentage of total speaking time), speech rate, and mean length of run. The number of unfilled pauses was calculated by counting the number of pauses of 1 second or more that occurred in test-takers' speech (Mehnert, 1998). Repair involved repetition of exact words, syllables, or phrases; replacements; reformulations (grammatical correction of structural features); false starts; and the partial repetition of part of a word or utterance (Freed, 2000). Total pause time was calculated by adding up all of the pauses. In order to enable comparisons across tasks (each of which allowed different amounts of time for completion), instances of filled pauses, unfilled pauses, and repair were counted per 60 seconds of speech. Speech rate was calculated by dividing the total number of syllables produced in a given speech sample by the amount of total time, expressed in seconds (Ortega, 1999). First, the transcribed speech was pruned by excluding features of repair; then the resulting total number of syllables was divided by the total speech time (excluding pauses of 3 or more seconds).[11] Mean length of run was calculated in terms of the mean number of syllables produced per utterance

(Towell, Hawkins, & Bazergui, 1996). (Note that mean length of run differs from mean length of utterance, which is a measure of the number of morphemes, not syllables, per utterance.)

Content

Two aspects of content were examined: quantity and quality. In order to measure quantity, counts of the number of T-units and clauses in the speech sample were used rather than the number of words, as the number and variety of words used is not a reliable guide to the number of ideas presented. The quality of the speech sample was examined by looking at the overall textual or schematic structure of the discourse, using the clause as the unit of analysis. Because of the time-consuming nature of the analysis, 40 performances on two tasks—20 for Task 2 (independent, music) and 20 for Task 3 (integrated, groundwater)—were chosen for analysis. For Task 2, a speaking task, structure was based on an analysis of test-takers' speech samples and a review of judges' comments. The analyses drew on terms and concepts used within the Sydney school of genre analysis (e.g., Eggins & Slade, 1997; Martin & Rothery, 1986; Paltridge, 2000). Central to the analysis of generic structure in this approach is the notion of obligatory and optional moves and the possibility of recursion of moves. For Task 3, the monologic listening-speaking task, in order to examine the overall textual structure of the speech, an ideal schematic structure was developed based on an analysis of the input text.

As Figures 1 and 2 show, the two tasks analyzed were found to have distinctively different schematic structures. The prompt for Task 2 is "What is your opinion about the value of art and music courses in a student's educational training?" As Figure 1 shows, the schematic structure needed to perform Task 2 allows an optional initial move: an introductory statement in which the topic (i.e., "whether art and music should be a part of school curriculum") is stated. Then follow the obligatory elements: the speaker's opinion on the topic, together with a reason or reasons for the opinion. A number of optional moves may then follow. For example, to clarify the point, one or more examples may be given. An alternative opinion may be presented, again followed by reasons and examples. Finally, the text may end with a concluding move in the form of a restatement of the speaker's opinion.

Task 3 requires test-takers to explain a series of problems and their solutions on the basis of input provided in a short lecture. The lecture presents the details of certain problems that occurred in the San Joaquin Valley, the causes of the problems, and the efforts that were made to solve them. In the lecture, a series of events in the valley is described chronologically. Analysis of the schematic structure for this task (see Figure 2) involves two levels, with second-level moves acting as a realization of each first-level move. The first level of structure involves four major obligatory moves (i.e., problem, solution, complication, and solution). At the second level there are three moves: process, outcome, and evaluation, the last of which is optional.

Note. Brackets [ ] denote an optional move.

Figure 1. **Schematic structure for Task 2.**

As Figure 2 shows, responses to Task 3 may commence with an optional introductory statement that describes the topic of the lecture. This is followed by a presentation of the problem, in which the cause of the problem (i.e., "process") is described, followed by its outcome and an optional evaluation of the outcome. A solution to the problem is then presented, which includes an explanation of what was carried out to solve the problem and its outcome; an optional evaluation of this outcome may be presented here. The next structural element at the first level is complication, in which the cause of a further problem, its outcome, and an optional evaluation are related, leading to some sort of crisis. This is followed by a second solution stage in which a description of another solution, its outcome, and an optional evaluation are presented. The text may end with an optional concluding statement.

Level 1	Level 2
[Introduction]	
Problem	Process
	Outcome
	[Evaluation]
Solution	Process
	Outcome
	[Evaluation]
Complication	Process
	Outcome
	[Evaluation]
Solution	Process
	Outcome
	[Evaluation]
[Conclusion]	

Note. Brackets [ ] denote an optional move.

Figure 2. **Schematic structure for Task 3.**

Summary of Speech Samples Analyses

Table 13 summarizes the analyses carried out on the speech samples. It lists conceptual categories, units of analysis, tasks used for the analyses, and the types of statistical analyses employed. For most features, results are reported in terms of inferential statistics. The exceptions are intonation, rhythm, and quality of discourse; it was not possible to use inferential statistics in the case of intonation and rhythm because of the small number of observations and, in the case of quality of discourse, because of the qualitative nature of the analysis. Descriptive statistics only are provided for intonation and rhythm. For the inferential statistical analysis, Analysis of Variance (ANOVA) with two factors was used for most of the data. The design was a five-by-two two-way ANOVA with five score levels and two task types. Some of the ratio measures (type-token and the first three complexity measures) were also sensitive to the amount of speaking time allowed per task, and in these cases, instead of adjusting for this in the measure itself, Analysis of Covariance (ANCOVA) was performed in order to eliminate the potential effect of the amount of speech, which was entered as a covariate.[12]

For both the ANOVA and ANCOVA analyses, some data were skewed and variances were not homogenous. In these cases, in accord with the usually suggested remedy, transformation of the data (e.g., a log or square root transformation) was considered. However, after consultation with a statistician, it was decided not to use transformed data for two reasons: (a) transformed variables are generally hard to interpret, and (b) both ANOVA and ANCOVA statistics are robust to violations of their assumptions, especially when the sample size is large.

For most measures, effect size is reported using *eta*. A strength of association measure such as effect size assesses the importance of any significant findings. Effect sizes may be interpreted as follows: *eta* < 0.2 = marginal; *eta* > 0.2, < 0.5 = small; *eta* > 0.5, < 0.8 = medium; *eta* > 0.8 = large.

Intercoder reliability. A number of means were adopted to ensure adequate intercoder reliability, appropriate to the kind of coding being attempted. The phonological analysis was carried out by two trained phoneticians., who used the following procedure to establish adequately high intercoder agreement. In the earlier stages, while they were refining the feature categories to use, the two coders went through data from 10 test-takers' performances together. In the test-run for the final categories, they first went through three test-takers' responses together, then transcribed five performances independently. Finally, they compared their results for the data from the five test-takers. In comparing counts of nontarget features, the two coders differed in the following way: Over the five responses transcribed independently, their feature counts differed on average by at most one token for both marginally nontarget-like or clearly nontarget-like features.

For the assessment of other analytic categories, approximately 10% of the data was coded by a second coder to calculate intercoder reliability, which was calculated using the Spearman-Brown prophesy formula (Henning, 1987:83). Table 14 presents the results of these reliability analyses. The achieved levels were high in almost all cases, with the exception of two grammatical accuracy features (use of tense and third-person-singular verbs) and one of the logical connectives (and); in these cases, levels were marginally below 0.8.

Table 13

Summary of Speech Sample Analyses

Category	Unit of analysis	Tasks analyzed	Method of analysis
Linguistic resources: Grammatical accuracy			
Specific errors (i.e., article use, third-person-singular verb, tense-marking, plural nouns, prepositions)	Target-like-use analysis: Percentage of forms supplied in both obligatory and nonobligatory contexts that are used correctly	All 5 tasks	ANOVA
Global accuracy	Percentage of error-free T-units	All 5 tasks	ANOVA
Linguistic resources: Grammatical complexity			
Sentence complexity			
T-unit complexity ratio	Number of clauses per T-unit	All 5 tasks	ANCOVA
Dependent-clause ratio	Ratio of dependent clauses to total number of clauses	All 5 tasks	ANCOVA
Verb-phrase ratio	Number of verb phrases per T-unit	All 5 tasks	ANCOVA
Mean length of utterance	Number of morphemes per utterance	All 5 tasks	ANCOVA
Sophistication			
Modals	Number of modals per 100 words	All 5 tasks	ANOVA
Comparative forms	Number of comparative forms per 100 words	All 5 tasks	ANOVA
Passive forms	Number of passive forms per 100 words	All 5 tasks	ANOVA
Linguistics resources: Textualization			
Use of logical connectives	Number of logical connectives used	Task 2 (N = 30) Task 3 (N = 40)	ANOVA

(Table continues)

Table 13 (continued)

Category	Unit of analysis	Tasks analyzed	Method of analysis
Linguistics resources: Vocabulary			
Word-token	Number of word-tokens	All 5 tasks	ANOVA
Word-type (general)	Number of word-types	All 5 tasks	ANOVA
Type-token ratio		All 5 tasks	ANCOVA
Word-type: K1, K2, AWL, Offlist	Percentage of total words used that are from K1, K2, AWL, or off-list	All 5 tasks	ANOVA
Phonology			
Pronunciation	Word-level (per 10 words): Meaningful words, on target Meaningful words, marginally nontarget Meaningful words, clearly not-target-like Nonmeaningful words	Tasks 2 & 3 (30 s only)	ANOVA (1) Descriptive (2)-(4)
	Subword-level (per 10 syllables): Target-like syllables Marginally nontarget-like syllables Clearly nontarget-like syllables Epenthetic syllables, clear Epenthetic syllables, marginal	Tasks 2 & 3 (30 s only)	ANOVA (1) Descriptive (2)-(5)
Intonation	Classification as: English-like, near English-like, not English-like, or few	Tasks 2 & 3 (30 s only)	Descriptive
Rhythm	Classification as: Stress-timed, variable, unclear, syllable-timed	Tasks 2 & 3 (30 s only)	Descriptive
Fluency			
Filled pauses (um and ers)	Number of filled pauses per 60 s	All 5 tasks	ANOVA
Unfilled pauses	Number of unfilled pauses per 60 s	All 5 tasks	ANOVA
Total pause time	Pause time as a percentage of required speaking time	All 5 tasks	ANOVA
Repair	The number of repairs per 60 s	All 5 tasks	ANOVA
Speech rate	Total number of syllables divided by total speaking time	All 5 tasks	ANOVA

(Table continues)

Table 13 (continued)

Category	Unit of analysis	Tasks analyzed	Method of analysis
Mean length of run	Mean number of syllables per utterance	All 5 tasks	ANOVA
Content			
Quantity	Number of T-units and clauses	All 5 tasks	ANOVA
Quality	Schematic structure	Task 2 (N = 30) Task 3 (N = 40)	Descriptive

Note. K1 = the most frequent 1,000 words of English; K2 = the second most frequent 1,000 words of English; AWL = words from the Academic Word List; Off-list = all remaining words. Instances of word-types and word-tokens are converted into frequency data (per 60 s) when they are compared across tasks.

Table 14

Summary of Intercoder Reliability

Category	Coding units	*N*	Intercoder reliability
Linguistic resources: Grammatical accuracy			
Specific types of errors	*Correct*		
	Articles	19	0.96
	Plurals	19	0.98
	Prepositions	19	0.98
	Tense	19	0.99
	Third-person-singular verbs	19	0.91
	Error		
	Articles	19	0.93
	Plurals	19	0.86
	Prepositions	19	0.88
	Tense	19	0.71
	Third-person-singular verbs	19	0.78
Global accuracy	Error-free T-units	19	0.98
	T-units with errors	19	0.99

(Table continues)

Table 14 (continued)

Category	Coding units	*N*	Intercoder reliability
Linguistic resources: Grammatical complexity			
Sentence complexity	T-units	20	0.91
T-unit-complexity ratio	Clauses		0.94
		20	
	Dependent clauses	20	0.99
Dependent-clause ratio	Verb phrases	20	0.98
Verb-phrase ratio	Morphemes	20	0.99
Mean length of utterance	T-units		
Sophistication		20	0.97
	Modals	20	0.99
	Comparative forms	20	0.94
	Passive forms	20	0.91
Linguistic resources: Textualization			
Use of logical connectives	And (add/ext/para)	10	0.99
	And (add/int/para)	10	0.76
	Because (cons/ext/hypo)	10	0.99
	But (comp/ext/para)	10	0.83
	But (comp/int/para)	10	0.99
	So (cons/ext/para)	10	0.99
	Then (tem/ext/para)	10	0.94
	When (tem/ext/hypo)	10	0.99
Fluency			
Filled pauses (um and ers)	Filled pauses	20	0.99
Repair	Repairs	20	0.98
Mean length of run	Syllables	20	0.99
	Utterances	20	0.98
Content			
Quantity (T-units and clauses)	See above		

Note. add/ext/para = additive external paratactic; comp/ext/para = comparative external paratactic; temp/ext/para = temporal external paratactic; cons/ext/para = consequential external paratactic; add/int/para = additive internal paratactic; comp/int/para = comparative internal paratactic.

Overall Results

As many and rather complex analyses of multiple variables were conducted for Study II, when reading this section it may help the reader to remember that we were interested in the extent to which the speech data varied according to (a) level of proficiency and (b) task type (independent v. integrated). Thus, two sets of results are reported for each measure: variation by level (Research Question 4) and variation by task type (Research Question 5). We present all the results by level first, and then report the results by task type.

As the results show, some of the measures proved to be a little disappointing; due to small observed frequencies in these instances, no sensible conclusions could be drawn. In retrospect, because of the delicacy of some measures, data and time in excess of the project's resources would have been required to yield meaningful results for them. In this section we necessarily report results for all of the measures used, but the reader is encouraged to note particularly results that are statistically significant as well as those with other than marginal effect sizes.

Results: Research Question 4—Comparison Across Levels

Research Question 4 seeks ways to characterize increasing levels of proficiency and examines the extent to which the findings of the discourse analysis confirm judges' perceptions proficiency. The analyses that pertain to these issues are discussed by conceptual category in the following paragraphs.

Linguistic Resources

Grammatical accuracy. Figures 3 and 4 graphically present the results for the various measures of grammatical accuracy (see Appendix E, Tables E1 and E2, for means and *SD*s). For all six measures Levels 4 and 5 are distinct, but the pattern is less clear at the lower levels. Five-by-two ANOVA analyses were performed separately for each variable, with target-like-use as the dependent variable for article use, tense-marking, use of third-person singular, plurals, and prepositions, and percentage of error-free T-units (global accuracy). For each variable, highly significant differences were observed (articles: F [4, 188] = 3.41, p = 0.001, *eta* = 0.07; tense-marking: F [4, 175] = 7.45, p = 0.001, *eta* = 0.15; third-person-singular: F [4, 139] = 3.01, p = 0.02, *eta* = 0.08; plurals: F [4, 173] = 9.58, p = 0.001, *eta* = 0.17; prepositions: F [4, 188] = 7.42, p = 0.001, *eta* = 0.14; global accuracy: F [4, 188] = 13.51, p = 0.001, *eta* = 0.22). No significant interactions were found. Note that the effect sizes (*eta*) were all marginal for the specific error

measures; as we might expect, a somewhat larger effect size (still small) was found for global accuracy (see Appendix G, Table G1, for detailed results).

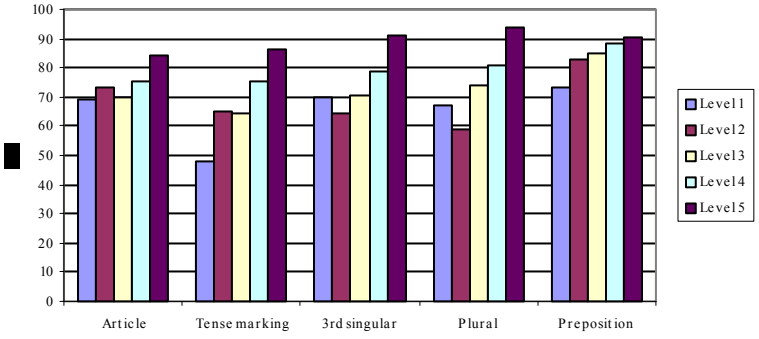


Figure 3. **Specific grammatical errors by proficiency level.**

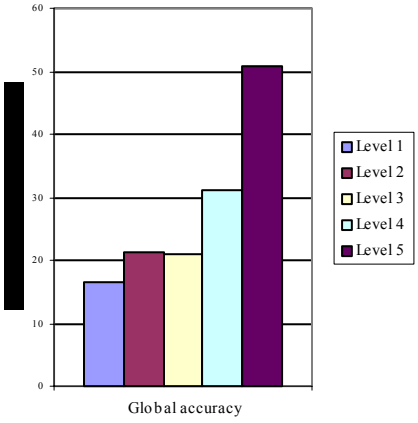
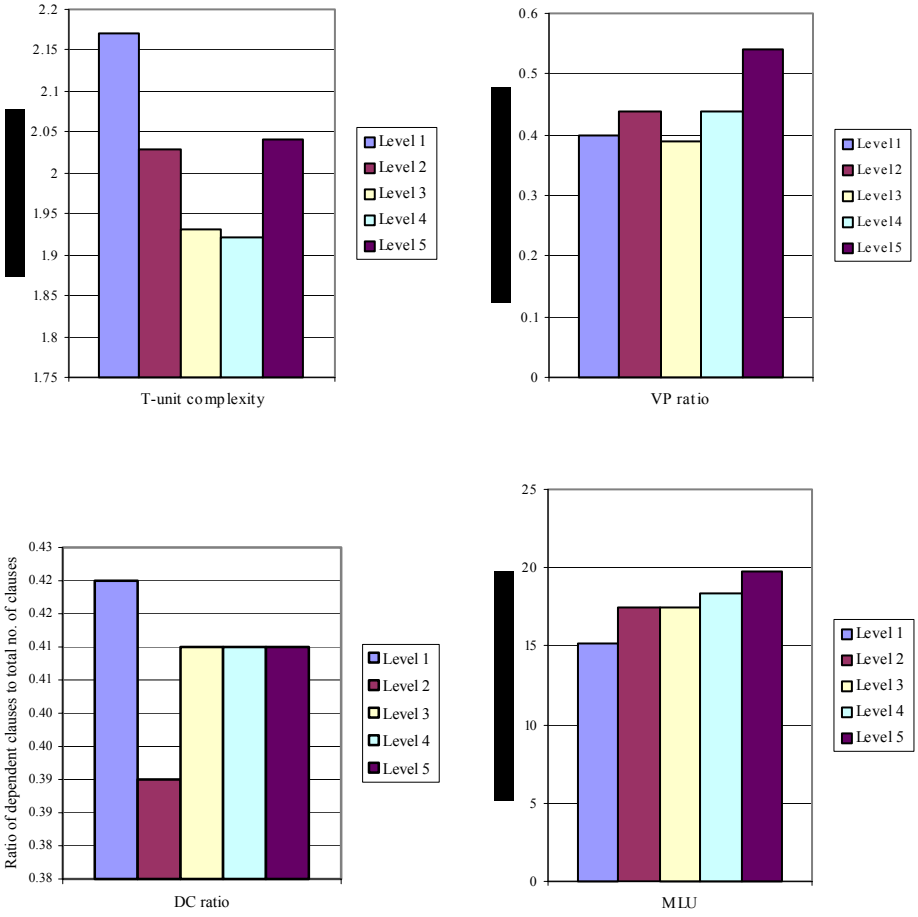


Figure 4. **Global accuracy by proficiency level.**

Grammatical complexity. Figure 5 presents the results by proficiency level for each of the two measures used to examine grammatical complexity: sentence complexity and sophistication. The results were somewhat mixed. The expected gradient of increasing complexity per level was observed for one of the measures, mean length of utterance. No such pattern was observed for T-unit-complexity ratio, dependent-clause ratio, or verb-phrase ratio (see Appendix E, Table E3, for descriptive statistics of each measure). However, when the number of utterances produced in each performance was taken into consideration in an analysis of covariance (ANCOVA[13]), significant differences were found across levels for verb-phrase complexity (F [4, 182] = 3.50,

$p = 0.01$, *eta* = 0.07), in addition to mean length of utterance (F [4, 187] = 10.77, $p = 0.001$, *eta* = 0.19); both effect sizes were marginal (see Appendix G, Table G2, for detailed results).



Note. VP ratio = verb-phrase ratio; DC ratio = dependent-clause ratio; MLU = mean length of utterance.

Figure 5. **Sentence complexity measures by proficiency level.**

Sophistication was measured by the frequency test-takers' use of comparatives, passive forms, and modals per 100 words. The measures were unrevealing, mainly because very few passive and comparative forms were observed across all levels (Figure 6), and as a result, little patterning emerged. (See Appendix E, Table E4, for means and *SD*s and Appendix G, Table G3, for ANOVA results. The dependent variable for each ANOVA analysis was the frequency of each structure per 100 words.)

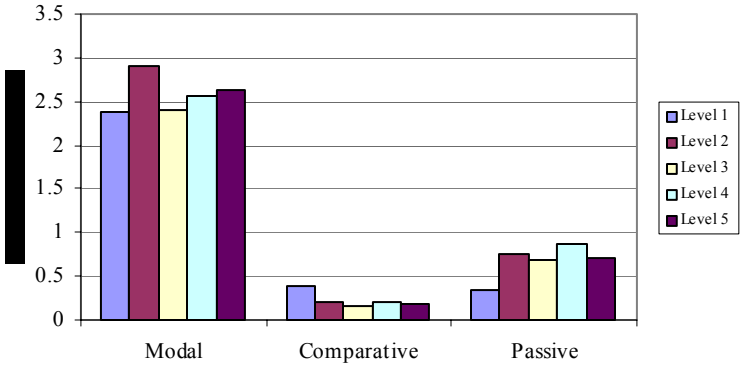
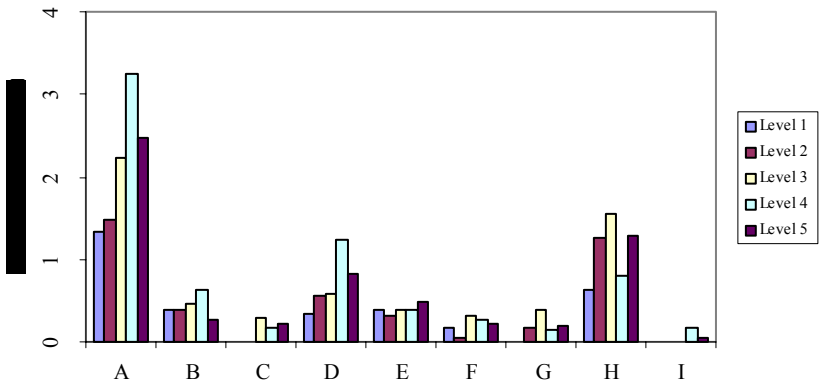


Figure 6. **Grammatical sophistication measures by proficiency level.**

Textualization. Figure 7 summarizes test-takers' use of the nine types of logical connectives studied. In general, the frequency of connectives per 100 words was very low across levels, and in some cases, no connectives were used at all (see Appendix E, Tables E5 and E6, for detailed results). No statistical analyses were performed, as the instances of each connective were very small.



Note. A = additive/external/paratactic "and;" B = comparative/external/paratactic "but;" C = temporal/external/paratactic "then;" D = consequential/external/paratactic "so;" E = additive/internal/paratactic "and;" F = comparative/internal/paratactic "but;" G = temporal/external/hypotactic "when;" H = consequential/external/hypotactic "because;" I = comparative/internal/hypotactic "not only…but also."

Figure 7. **Use of logical connectives by proficiency level.**

Vocabulary. Figure 8 displays results for the word-token, word-type, and type-token ratio vocabulary measures,. As Figure 8 shows, increases in proficiency were associated with an increase in the number of words produced (word-tokens) and with use of a wider range of words (word-type). ANOVA analyses (with the number of word-tokens, the number of word-types, and the type-token ratio as dependent variables) confirmed significant differences for word-token and word-type (word-token: F [4, 190] = 62.32, p = 0.001, *eta* = 0.57; word-type: F [4, 190] = 47.88, p = 0.001, *eta* = 0.50) with medium effect sizes (*eta* = 0.57 and 0.50, respectively; see Appendix E, Table E7, for descriptive statistics and Appendix G, Table G4, for detailed statistical results).

Counterintuitively, type-token ratios were larger for lower-level learners than higher-level learners. In order to adjust for any effect of length of speech, the number of utterances produced in each performance was taken into account in an ANCOVA analysis; the results still confirmed a significant effect for level in this unexpected direction (F [4, 189] = 13.23, p = 0.001, *eta* = 0.22; see Appendix G, Table G4, for detailed results).

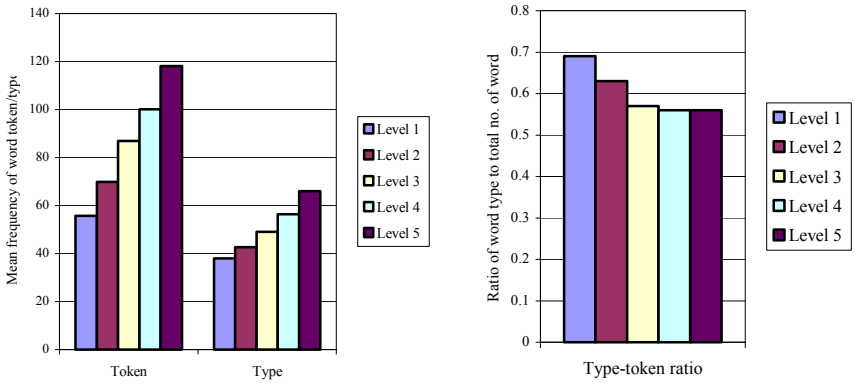
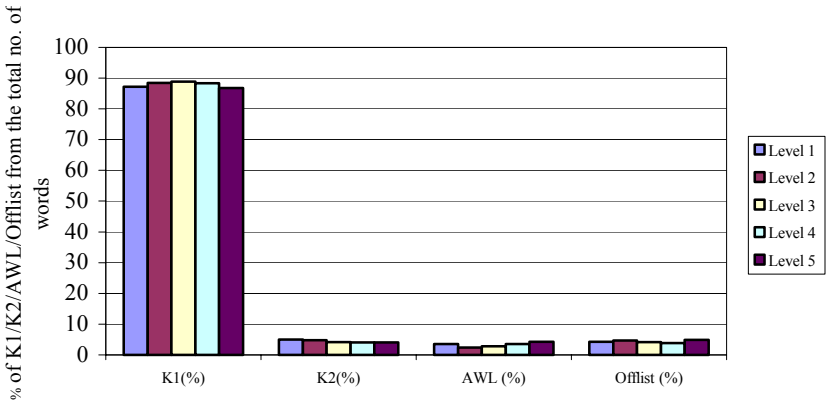


Figure 8. **Vocabulary measures (1) by proficiency level.**

Figure 9 depicts results for word use according to word list frequency categories. When we examined the proportions of word-types falling into different frequency categories, we expected that the percentages of the words belonging to the K2 and AWL categories would be sensitive to level. This proved not to be the case. As Figure 9 shows, the majority of the words test-takers used were from K1, and only a very small portion of words belonged to AWL or K2. ANOVAs with dependent variables being the percentage of K1, K2, AWL, and off-list words used revealed a

significant difference across levels only for the percentage of AWL words used (F [4, 190] = 3.99, p = 0.001). The effect size was very small (*eta* = 0.08; see Appendix E, Table E8, for descriptive statistics and Appendix G, Table G5, for detailed results). Significant interactions were found for the percentage of K1 and K2 words used, which makes the interpretation of main effects a less than straightforward endeavor in this case. (No other significant interactions were found in the entire study.)



Note. K1 = the most frequent 1,000 words of English; K2 = the second most frequent 1,000 words of English; AWL = words from the Academic Word List; Off-list = all remaining words.

Figure 9. **Vocabulary measures (2) by proficiency level.**

Phonology. Several of the features of phonology that were analyzed proved sensitive to differences in proficiency level. The various subcategories are presented in turn. Figures 10 and 11 present the results for pronunciation. As Figure 10 shows, the proportion of meaningful words classified as showing target-like pronunciation increased across levels, with the exception of Level 1 (see Appendix E, Table E9, for detailed results). The proportion of words whose pronunciation was classified as marginally nontarget or clearly nontarget was not sensitive to level; nor was the number of words classified as non-meaningful. However, when an ANOVA analysis was performed to compare the frequency of meaningful words that were on target across levels (with the number of meaningful words on target per 10 words as the dependent variable), no significant difference was observed (see Appendix G, Table G6, for detailed results). No analysis was

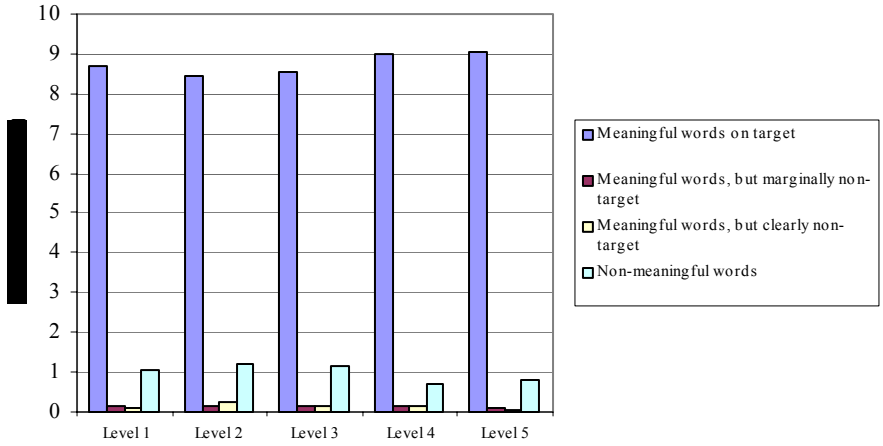


Figure 10. **Word pronunciation by proficiency level.**

performed on the other categories, as frequencies were very low.

As Figure 11 shows, more noticeable differences across levels were found in the subword-level analysis (see Appendix E, Table E9, for detailed results). In general, the number of nontarget-like syllables (especially marginally nontarget-like syllables) was sensitive to proficiency level (see Appendix G, Table G6). The results of an ANOVA with the number of target-like syllables per 10 syllables as the dependent variable showed a highly significant difference across levels (F [4, 69] = 11.49, p = 0.001); the effect size was small (*eta* = 0.40). Again, statistical analyses of the other categories were not seen as meaningful because of the low observed frequencies associated with them.

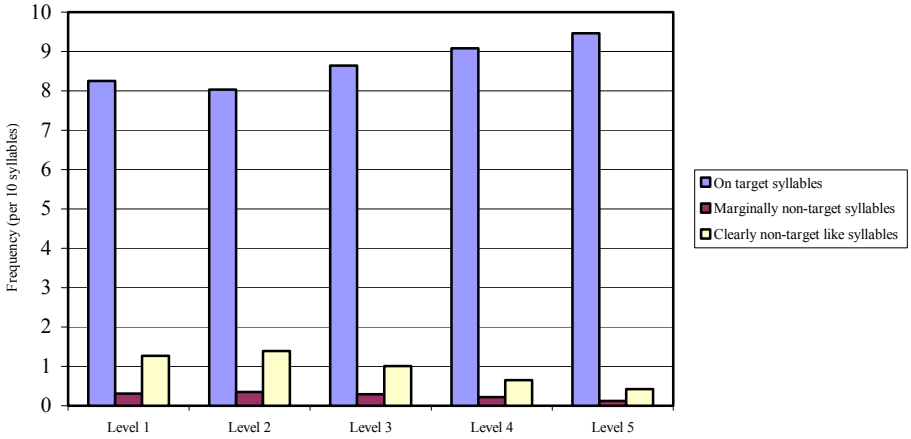
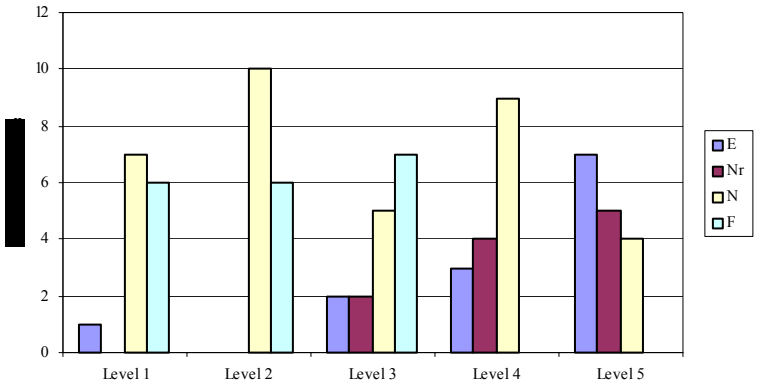


Figure 11. **Syllable pronunciation by proficiency level.**

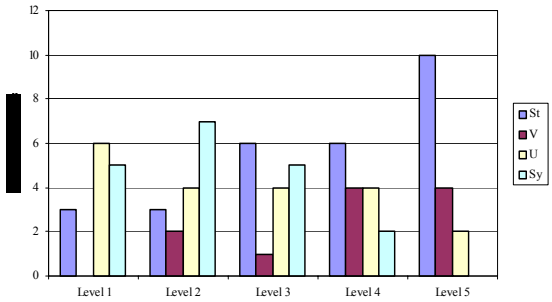
Intonation. The expected pattern emerged in the intonation analysis: The intonation units of higher-level learners were more frequently categorized as many and English like, compared with those of lower-level learners. More than half of the lower-level test-takers fell into the two lowest performance categories ("many and not English-like" and "few"). Few lower-level test-takers achieved English-like intonation, and many were assessed as performing in the not-English-like category. As Figure 12 shows, half of all test-takers below Level 3 fell into the two lowest categories (see Appendix E, Table E10, for detailed results). No statistical analysis was performed for intonation due to the small frequencies in each category.



Note. E = many & English-like; Nr = many & near English-like; N = many & not English-like; F = Few.

Figure 12. **Intonation measures by proficiency level.**

Rhythm. Measures of rhythm (or word stress) also proved sensitive to differences in level: Few lower-level speakers were assessed as managing stress-timed speech, which seemed to be a characteristic of higher-level performances. The rhythm of more than half of the Level 4 and 5 test-takers was coded as stress-timed; in contrast, the coders' judgments of the rhythm of the speech of many Level 1 and 2 test-takers fell in the categories unclear and syllable-timed (see Figure 13). Level 3 and Level 4 test-takers were indistinguishable in terms of the numbers who were assessed as stress timed, but more Level 3 test-takers were assessed as unclear and syllable-timed than Level 4 test-takers (see Appendix E, Table E11, for detailed results). No statistical analysis was performed for these intonation measures due to the small frequencies in each category.



Note. St = stress-timed; V = variable; U = unclear; Sy = syllable-timed.

Figure 13. **Rhythm measures by proficiency level.**

Fluency. Figure 14 displays the results for the various measures of fluency. The results for speech rate, number of unfilled pauses, and total pause time showed a clear relationship with proficiency level. Higher level test-takers spoke faster, with less pausing, and with fewer unfilled pauses than lower-level test-takers (see Appendix E, Table E12, for detailed results).

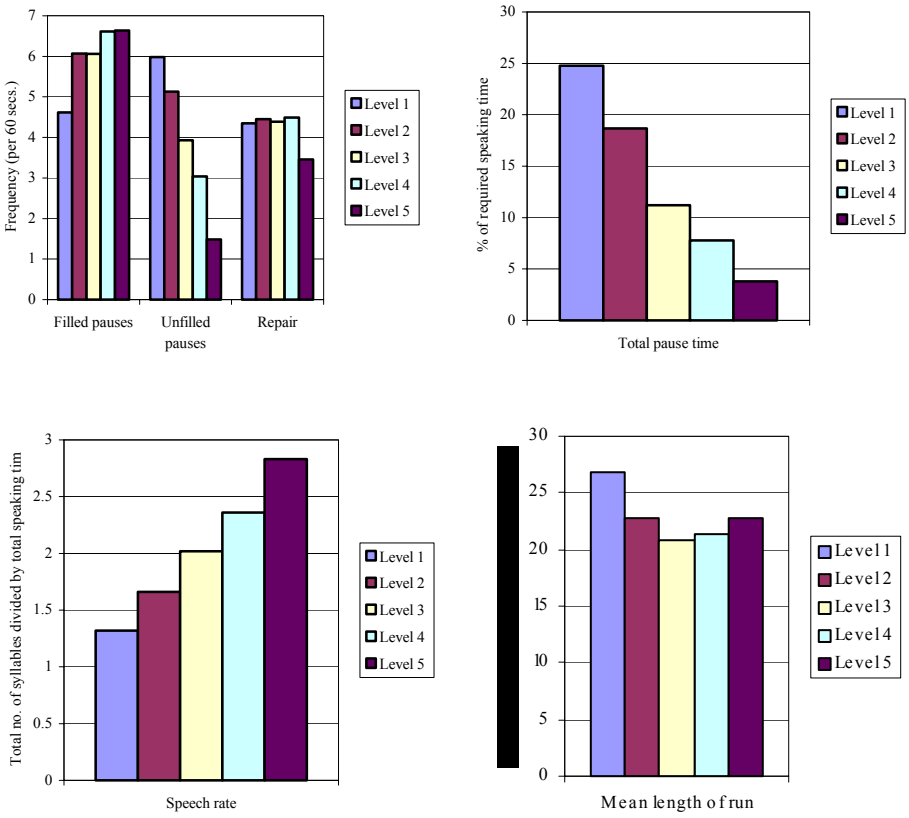


Figure 14. **Fluency measures by proficiency level.**

ANOVAs were carried out with the number of filled pauses per 60 seconds, the number of unfilled pauses per 60 seconds, total pause time (as a percentage of total speaking time), the number of repairs per 60 seconds, speech rate (total syllables per speaking time), and mean length of run (syllables per utterance) as dependent variables. Significant differences were found for speech rate (F [4, 189] = 71.32, p = 0.001, *eta* = 0.60); unfilled pauses (F [4, 190] = 12.19, p = 0.001, *eta* = 0.20), and total pause time (F [4, 190] = 20.62, p = 0.001, *eta* = 0.30), with medium or small effect sizes (see Appendix G, Table G7, for detailed results).

Content. Of the two types of measures used to examine the content of test-takers' performances, quantity and quality, the *quantity* measures yielded mixed results. As Figure 15 shows, the number of T-units was not uniformly different across proficiency levels, but the number of clauses increased gradually as proficiency level went up (see Appendix E, Table E13, for descriptive statistics). Five-by-two ANCOVA analyses with the number of T-units per 10 utterances and the number of clauses per 10 utterances as dependent variables showed highly significant differences for both T-units (F [4, 187] = 17.42, p = 0.001) and clauses (F [4, 187] = 28.76, p = 0.001). The effect sizes (*eta*) in each case were small (0.27 and 0.38, respectively). ANCOVA statistics were used to eliminate the effect of the amount of speech (i.e., with the number of utterances as covariate; see Appendix G, Table G8, for detailed results).

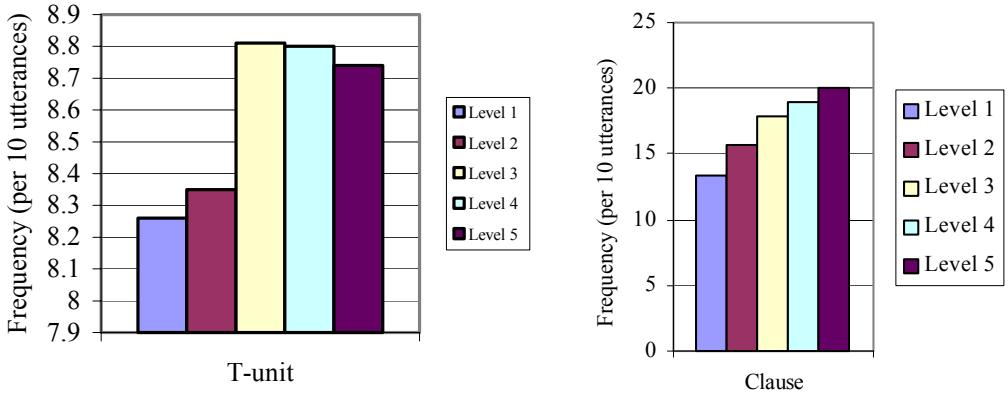


Figure 15. **Quantity of discourse by proficiency level.**

The quality measures showed a clear relationship to proficiency level. In terms of the schematic structure developed for Task 2 (shown previously in Figure 1), lower-level test-takers gave their opinion, but generally did not provide reasons. This is unsurprising given that the length of their performances was generally very short (i.e., one or two utterances). In Example 1, which follows, the test-taker starts his/her speech with a statement of opinion (i.e., opposing the cancellation of classes) and goes on to provide one brief reason for the opinion, without any further moves.

Example 1—Task 2, Proficiency Level 1

Structure	Test-taker's speech
Opinion	I think it's not a good idea to use the #2 budget by eh #2 cancel classes like uh music and art courses.
Reason	Um, #2 I think for a good training it's um #5 it's better to have um #3 art and musical courses #6 mm #9

More elaborate structures were observed in the performances of Level 3 speakers, but their speech was often repetitive and not very clear. For instance, in Example 2, the speaker commences with a statement of opinion, followed by a number of reasons for the opinion. However, the speaker does not provide examples to illustrate his/her disagreement with the claim made in the newspaper.

Example 2—Task 2, Proficiency Level 3

Structure	Test-taker's speech
Opinion	I disagree with the um with the point of views shown in the newspaper.
Reason	I think um um everyone has different interest
Reason	and um they ap- um everyone is #1 good at different things.
Reason	And if um so if the um school um #2 cancelled um music and art courses #1 um um, it's too cruel to the students who like um music and art.
Reason	Um, since many people ums ability um uh are #2 are built up in um in school
Reason	if um they cannot receive the education of their um interest they will lose the um chance to um be able to do it best.

In contrast, more elaborate structures including examples and clear illustrations of points were found in performances by Level 4 and 5 speakers (Example 3).

Example 3—Task 2, Proficiency Level 5

Structure	Test-taker's speech
Reason	Training in art music is indispensable to the overall growth of a student.
Opinion	Therefore I think #1 it is very unfortunate that music and art courses might be cancelled because of budgetary constraints.
Example	The appreciation of music is something that one must learn as a child and leads to a whole lot of benefits over one's lifetime.
Reason	Similarly the appreciation of art is something which increases your perception
Example	and uh #2 may be of great importance to you if you become either an engineer or an architect to have greater perception and visualization skills.
Example	Similarly, music is what in some sense differentiates man from beast.
Restatement of opinion	It cannot be removed from one's learning.

In terms of the schematic structure developed for Task 3 (shown previously in Figure 2), performances by lower-level test-takers were very short, consisting of only one or two points and little structure, much like Task 2 performances. In Example 4, the Level 1 speaker commences with a statement of the problem, which is followed by an illustration and a comment on the problem. No other Level 1 structure is given apart from a statement of the problem.

Example 4—Task 3, Proficiency Level 1

Level 1	Level 2	Test-taker's speech
Problem	Process	I think the problem #2 is occurred in California #2 is the same #1 kind #1 very.
	Process	People #3 haven't enough #2 water to use.
	Evaluation	#4 And uh, #5 we will #6 we, #2 we can, we can live without the water.

As in Task 2, the schematic structure of higher-level performances was more elaborate than that of lower-level speakers for Task 3. For instance, Example 5, a response from a Level 3 candidate, comprises three Level 1 structures. A statement of the problem is illustrated well with chronological details and other relevant information. A brief explanation of a solution to the problem is presented, but the outcome of the solution is not mentioned (i.e., whether the land stopped sinking or not). Then in the next stage, a complication is well described, with the test-taker illustrating how the problem of land subsidence was caused and what happened then. However, the speech stops there, and no solution to the complication is provided.

Example 5—Task 3, Proficiency Level 3

Level 1	Level 2	Test-taker's speech
Problem	Process	The problem that occur in California #1 come from #2 um #2 the people who live over there um use a lot of water #2 that come from the underground water #3 and #2 make a lot of problem,
	Process	especially in nineteen seventy the water level drop
	Outcome	and #2 in the vast area #2 sank so much #1
Solution	Process	and they tried to solve the problem to decrease the amount of the water #1 or to buy the surface water #2
Complication	Process	but the problem still occur #3 because of the area hit with the drought #2
	Process	and the people still start pumping pumping the water again.
	Outcome	That is the main cause #2 of the problem that we can not absolutely solve the problems.
	Outcome	This problems that damage the surface of the land and cause to the ground to sink and create the environment problems and con #1 taminate problems and also increase, the pressure increase and ground problem.

Example 6, from a Level 5 candidate, contains more sophisticated structures than Example 5: All four Level 1 structures are present, each with two Level 2 structures. All crucial points presented in the speech are well illustrated and logically connected. There are no ambiguous statements.

Example 6—Task 3, Proficiency Level 5

Level 1	Level 2	Test-taker's speech
Problem	Process	Um, the problem that occurred there was due to um #1 the land, the over #2 over consumption of underground water.
	Process	Um, too too big amounts of water was being pumped up from the underground
	Outcome	and um this caused uh land subsidence uh #2 and subs-subsidence of the valley in the nineteen twenties.
	Outcome	And uh even though this this uh this problem occurred they continued to pumped a lot of water out of the ground which like just make the water level are really #1 sinking down much faster,
Solution	Process	so finally in the nineteen seventies they tried to stop the the process of of land subsiding by reducing the amount of water being pumped up from the from the ground, out of the groundwater.
	Process	And uh they decided to import water, surface water, from outside.
	Outcome	And um, the land recovered, the land subsidence stopped #2 um considerably
	Evaluation	and that was really good,
Complication	Process	but then in nineteen ninety the uh they were faced with a with a drought
	Outcome	which which didn't make their surface water available anymore
Solution	Process	so they had to pump the water again from the groundwater
	Process	and they did that and uh
	Outcome	but this time the water level uh sank much faster, like after that the valley has been affected
	Evaluation	and um #1 the problem were even worse.

Summary of Results for Research Question 4

Overall, a number of measures provided evidence that features of test-takers' discourse varied according to their assessed proficiency levels. Significant differences across proficiency levels in the expected directions were found for the following features:

- grammatical accuracy: article use, tense-marking, third-person-singular verbs, plural nouns, prepositions, and global accuracy
- grammatical complexity: verb-phrase ratio and mean length of utterance
- vocabulary: word-token, word-type, type-token ratio, and percentage of words from the Academic Word List
- pronunciation: target-like syllables
- fluency: speech rate, number of unfilled pauses and total pause time
- content: number of T-units and number of clauses

In addition, the quality measure for the content category showed clear proficiency level differences; and the measures of intonation and rhythm (aspects of phonology) appeared to be sensitive to proficiency level differences, although no statistical testing was possible. Table 15 provides a summary of the findings of the statistical analyses related to Research Question 4.

Table 15

Summary of Statistical Analyses by Proficiency Level

		Level	
		Difference	Effect size
Linguistic resources:	Article	√	0.07
Grammatical accuracy	Tense-marking	√	0.15
	Third-person-singular	√	0.08
	Plural	√	0.17
	Preposition	√	0.14
	Global accuracy	√	0.22
Linguistic resources:	T-unit complexity		
Grammatical complexity	Dependent-clause ratio		
	Verb-phrase ratio	√	0.07
	Mean length of utterance	√	0.19
	Modal		
	Comparative		

(Table continues)

Table 15 (continued)

		Level	
		Difference	Effect size
	Passive		
Linguistic resources:	Word-token	√	0.57
Vocabulary	Word-type	√	0.50
	Type-token ratio	√	0.22
	K1 (%)		
	K2 (%)		
	AWL (%)	√	0.08
	Off-list (%)		
Phonology: Pronunciation	Meaningful words		
	Target-like syllables	√	0.40
Fluency	Number of filled pauses		
	Number of unfilled pauses	√	0.20
	Total pause time	√	0.30
	Repair		
	Speech rate	√	0.60
	Mean length of run		
Content	T-units	√	0.27
	Clauses	√	0.38

Note. (a) √ = statistical difference; (b) Effect size (*eta*): marginal = (< 0.2); small (> 0.2 < 0.5); medium (> 0.5 < 0.8); large (> 0.8). (c) No statistical analyses were performed for use of logical connectives (textualization, subcategory of linguistic resources), intonation (phonology category), or rhythm (phonology category) due to the small number of instances in each category. (d) K1 = the most frequent 1,000 words of English; K2 = the second most frequent 1,000 words of English; AWL = words from the Academic Word List; Off-list = all remaining words.

While for all these features the results of the statistical analyses (ANOVA or ANCOVA) showed highly significant differences across levels, the effect sizes varied, with most being small or marginal.[14] In other words, for each of these variables, while the differences across level were real—that is, not attributable to chance—any one taken in isolation was not particularly strong in determining the overall score for the speaker. Table 16 lists those with the strongest impact on

Table 16

Relative Impact of Various Discourse Features on Scores

Feature	Feature category	Effect size	Effect size category
Speech rate	Fluency	0.60	Medium
Word-token	Vocabulary	0.57	Medium
Word-type	Vocabulary	0.50	Medium
Target-like syllables	Phonology: Pronunciation	0.40	Small
Number of clauses	Content: Quantity	0.38	Small
Total pause time	Fluency	0.30	Small
Number of T-units	Content: Quantity	0.27	Small
Global accuracy	Linguistic resources: Grammatical accuracy	0.22	Small
Type-token ratio[a]	Vocabulary	0.22	Small
Number of unfilled pauses	Fluency	0.20	Small

Note. Effect size (*eta*): small = (> 0.2 < 0.5); medium = (> 0.5 < 0.8).

[a] This was not in the expected direction (commentary follows).

overall scores in order.

The effect for quality (a measure of content) and the suggestive findings for intonation and rhythm (measures of phonology) must be added to this list. (In these cases it was not possible to calculate effect size.) In addition, significant effects with marginal effect sizes were found for a number of other variables: each of the specific measures of grammatical accuracy; two of the measures of grammatical complexity (verb-phrase ratio and mean length of utterance); and one vocabulary measure (percentage of words from the Academic Word List).

These findings are very interesting for a number of reasons. First, they reveal that features drawn from a wide range of categories were making independent contributions to the overall impression of the candidate. This is encouraging from the point of view of validity and appears to contradict the popular belief that perceptions of grammatical accuracy are the main drivers of oral proficiency scores (see McNamara, 1990). In the current study, features from each of the conceptual categories were among those having the greatest influence on overall scores; no category was omitted. The most diffuse but persistent influence appeared to be that of grammatical

accuracy: Each of the features in this category had a significant relationship to score levels. Also notably, more macrolevel measures—including speech rate, the main vocabulary measures, measures of quantity of ideas, a global pronunciation measure, and the global grammatical accuracy measure—appear to have most influence on scores, which is what we might expect.

Second, the results allow us to draw conclusions about the methodology used in the study. Not all of the measures proved to be useful. Sometimes, as we have noted, this was because the measures used were too delicate to yield results in a study of this size. Specifically, problems arose with the measures used for the sentence complexity and vocabulary subcategories. For sentence complexity, we found a problem with the ratio measures (the T-unit-complexity ratio, the dependent-clause ratio, and the verb-phrase ratio), even though, as stated above, these had been recommended on the basis of previous studies as among the most useful measures of complexity (Wolfe-Quintero et al., 1998).

If we look at a sample of actual performances at different proficiency levels (see Appendix H, Examples H1-H5), we note interesting apparent differences. First, is the sheer volume of clauses and T-units produced at the higher levels, which contrasts with the lower levels and was reflected in the content measures and two vocabulary measures (word-type and word-token). Second, there is some evidence of increasing complexity per level, though this is not a strong or uniform effect. Shorter and simpler sentences with little subordination were more frequently observed at lower levels, whereas complex sentences with several instances of subordination were generally a feature at higher levels, though there was not a strong difference between Levels 3, 4, and 5. The ratio measures reflected these differences relatively weakly. It is possible that these measures are useful only with longer samples of text, as they have previously been used mainly in the analysis of written discourse, and it was on the basis of those studies that Wolfe-Quintero et al (1998) recommended their use.

A particular problem arose with the results for the type-token ratio (a vocabulary measure): Type-token ratios were larger for lower-level test-takers than higher-level ones despite the fact that a greater number of words and word-types were observed at higher levels than at lower levels. Comments made by judges provided some initial indication of why this was so: Some lower-level test-takers simply repeated the prompt or part of the stimulus text, which naturally led to higher quality output than would otherwise be expected of test-takers at this level.

Results: Research Question 5—Comparison Across Task Types

Research Question 5 explores the extent to which differences across tasks and task types—as attended to by judges and described in the task specifications—could be confirmed through an empirical analysis of test-taker discourse. Thus, in this section we report on the analyses of comparisons of various aspects of test-takers' speech across task types. As noted in the Methodology section for Study II, most of these analyses, with the exception of three, involved all tasks; texualization (linguistics resources), phonology, and content were each analyzed on the basis of only two tasks—one independent and one integrated.

In general, the greater complexity of integrated tasks in terms of content and organization led us to expect differences in the quality of the content and organization of performances across the two task types. Further than that, our expectations of the potential differences between performances on the two task types in terms of more specific features were somewhat unclear. On the one hand, because integrated tasks provide learners with language input, we expected that the better responses to these tasks would involve more complex or sophisticated language, in terms of vocabulary at least—and given the greater potential complexity of ideas to be communicated, in terms of grammatical complexity as well. On the other hand, we felt the greater cognitive demands of integrated tasks could have the opposite effect on markers of linguistic processing, as they would make it more difficult for speakers to manage the linguistic control needed to yield higher scores on measures of sophistication and complexity.

The literature on information processing approaches to tasks (e.g., Robinson, 1995, 1996, 2002; Skehan, 1996, 1998) is similarly ambiguous. On the one hand, following Skehan, the assumed higher cognitive load of these tasks should mean that fewer cognitive resources are available to manage aspects of linguistic processing, resulting in lower scores on these measures; on the other, following Robinson, the greater cognitive challenge may lead to heightened concentration, yielding generally better performances and resulting in higher scores on at least some of the features measured.

In the remainder of this section, we report the results for each measure in some detail and discuss the implications of our findings. The reader who wishes to get a quick overview of the results should consult Table 17 before continuing.

Linguistic Resources

Grammatical accuracy. Grammatical accuracy was examined in terms of both specific types of errors and global accuracy. Figures 16 and 17 present a graphical summary of the results for each task (see Appendix F, Tables F1 and F2, for means and *SD*s for each task and task type).

The results of five-by-two ANOVA analyses with target-like-use analysis for articles, tense marking, third-person-singular verbs, plural nouns, and percentage of error-free T-units for global accuracy as dependent variables revealed only one significant difference across task types: Article use was more accurate in independent tasks than integrated tasks (F [1, 188] = 5.30, p = 0.02), but with a marginal effect size (*eta* = 0.03; see Appendix G, Table GI, for detailed results for each task type).

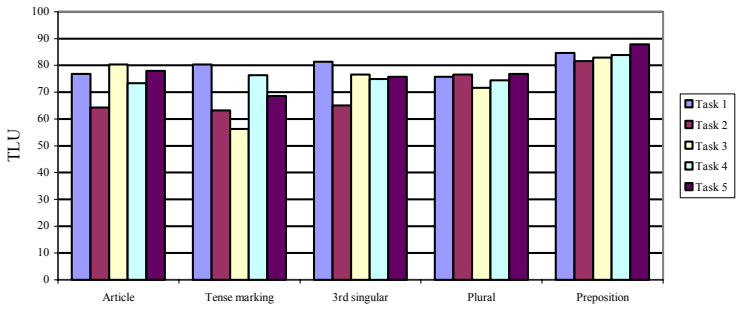


Figure 16. **Specific grammatical errors by task.**

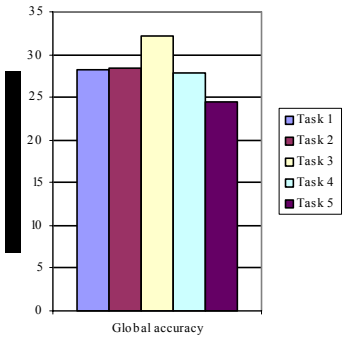
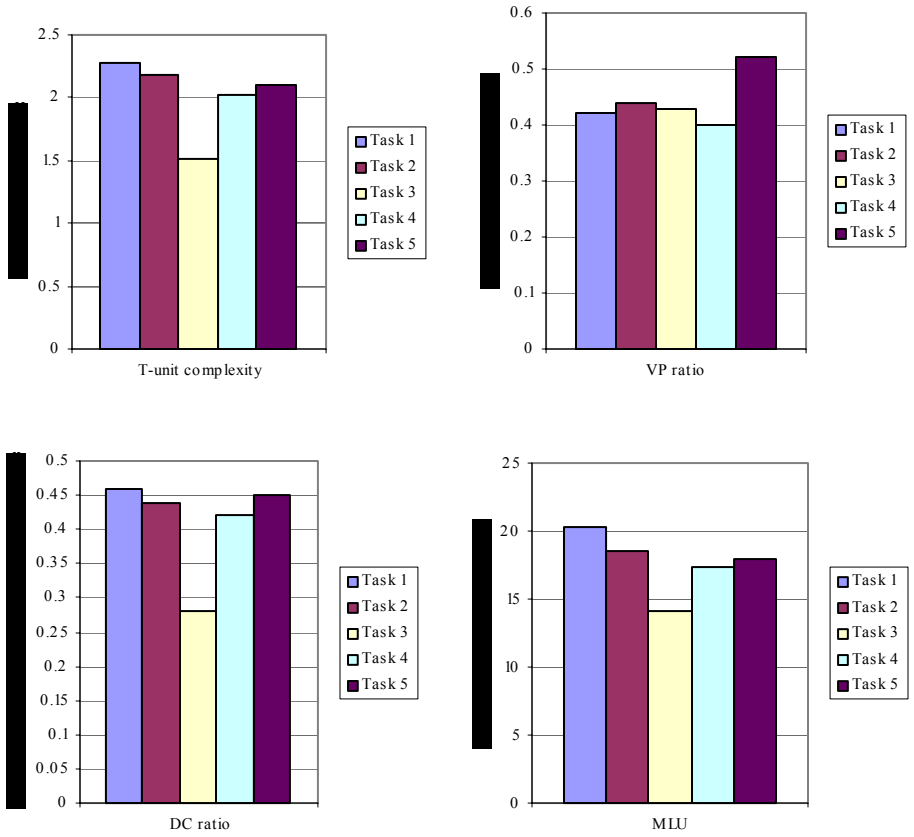


Figure 17. **Global accuracy by task.**

Grammatical complexity. Grammatical complexity was examined in terms of both sentence complexity and sophistication. Figure 18 presents the results for sentence complexity measures by task (see Appendix G, Table F3 for descriptive statistics). On two of the four measures (T-unit complexity and mean length of utterance) the means favored the independent tasks (Tasks 1 and 2); performance on Task 3 was by far the weakest. On the dependent-clause ratio, too, Task 3 (but this time alone) appeared markedly less complex. However, five-by-two ANOVAs with T-unit-complexity ratio, dependent-clause ratio, and mean length of utterance as dependent variables did not reveal any statistically significant differences (T-unit-complexity ratio: F [1, 179] = 1.54, p = 0.22; dependent-clause ratio: F [1, 181] =0.001, p = 0.97;mean length of utterance: F [1, 187]= 0.04, p = 0.84; see Appendix G, Table G2, for detailed results for each task type).

At first, there appeared to be no consistent differences across tasks for verb-phrase ratio. However, when the number of utterances produced in each performance was taken into consideration (the required length of speech was different from one task to another, and candidates' actual speaking time also varied), the adjusted means for verb-phrase ratio showed a difference favoring integrated tasks (see Appendix F, Table F6), a result confirmed by the five-by-two ANCOVA analysis with the number of utterances as the covariate (F [1, 182] = 4.11, p = 0.04, *eta* = 0.02). Again, however, the effect size (*eta*) was very marginal (see Appendix G, Table G2, for detailed results for each task type).

Figure 19 shows the results of the sophistication analysis (see Appendix F, Table F8, for descriptive statistics). Test-takers used more modals in independent tasks than integrated tasks, and when five-by-two ANOVAs were performed with the frequency of each structure per 100 words as dependent variables a significant difference was observed for this measure (F [1, 190] = 79.79, p = 0.001); the effect size was small (*eta* = 0.30). Test-takers used very few comparatives and passive forms across all tasks (see Appendix G, Table G3, for detailed results for each task type).



Note. VP ratio = verb-phrase ratio; DC ratio = dependent-clause ratio; MLU = mean length of utterance.

Figure 18. **Sentence complexity measures by task.**

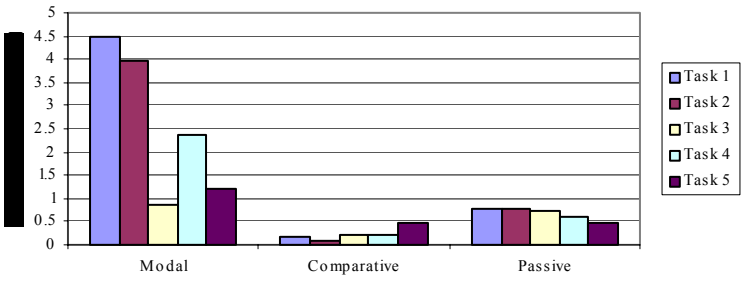
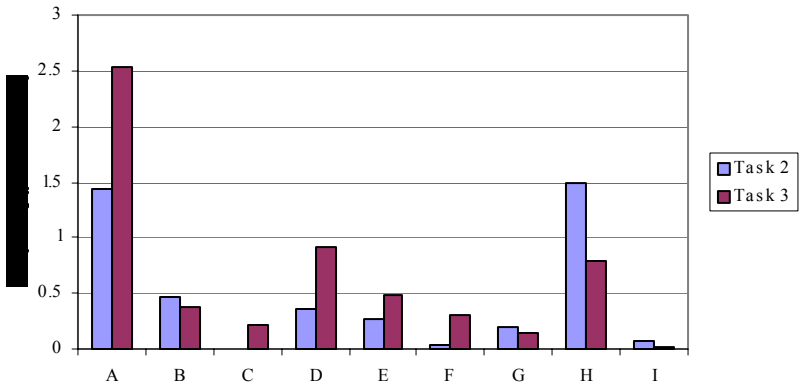


Figure 19. **Grammatical sophistication by task.**

Textualization. Figure 20 provides results for the frequency of logical connectives per 100 words in test-takers' responses to Tasks 2 (independent) and 3 (integrated). Few consistent differences were observed between the two tasks. Due to the low frequency of occurrence of individual connectives, no statistical analysis was performed (see Appendix F, Tables F9 and F10, for descriptive statistics for each task).



Note. A—additive external paratactic, "and"; B - comparative external paratactic, "but"; C - temporal external paratactic, "then"; D– consequential external paratactic, "so"; E - additive internal paratactic " and"; F - comparative internal paratactic "but"; G - temporal external hypotactic, "when"; H—consequential external hypotactic, "because" I—comparative internal hypotactic "not only…but also."

Figure 20. **Use of logical connectives by task.**

Vocabulary. ANOVAs were performed for word-type and word-token with the number of word-tokens and number of word-types as dependent variables (see Appendix F, Table F11, for descriptive statistics for each task and task type). As Figure 21 shows, both the word-token and word-type measures showed lower figures on the integrated tasks, particularly Tasks 3 and 5, which was confirmed by the results of five-by-two ANOVA analyses (Word-token: F [1, 190] = 25.22, p = 0.01; Word-type: F [1, 190] = 119.08, p = 0.001). The effect sizes (*eta*) were marginal (0.12) and small (0.39), respectively (see Appendix G, Table G4, for detailed results). We comment further on these findings in conclusion to this section, Summary of Results for Research Question 5.

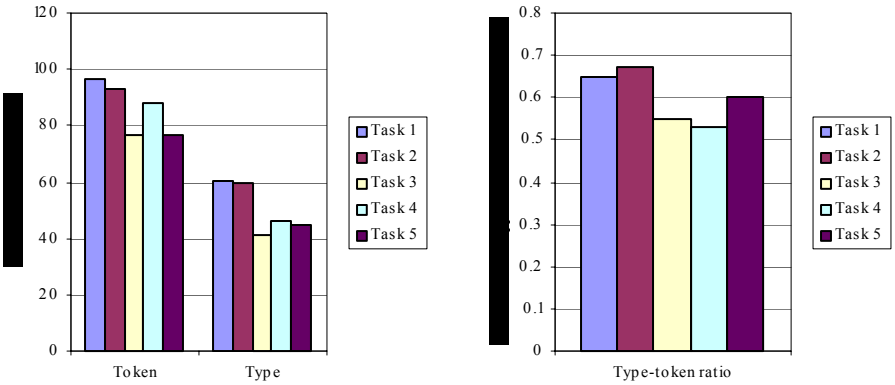
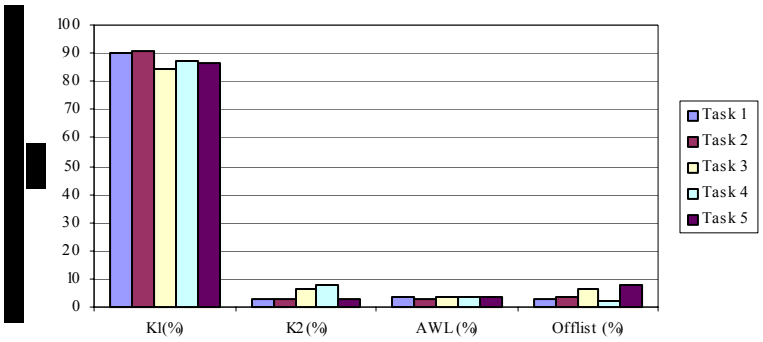


Figure 21. **Vocabulary use (1) by task.**

As Figure 21 also shows, we found a larger type-token ratio for the independent tasks (Tasks 1 and 2) than for the integrated tasks (Tasks 3-5; see Appendix F, Table F11). However, this effect was modified when the exact length of speech was taken into account in the analysis: The adjusted means that were calculated after the effect of length of speech was removed showed that the means of four of the tasks (i.e., the two independent tasks and two of the integrated tasks—Tasks 3 and 5) were nearly identical (see Appendix F, Table F12, for adjusted means and *SD*s.) The ANCOVA analysis (with type-token ratio as the dependent variable) still showed a significant difference (F [1, 189] = 4.42, p = 0.04), but the effect size was very marginal (*eta* = 0.02; see Appendix G, Table G4, for detailed results).

Figure 22 presents the results for the remaining vocabulary measures (those based classification by word list). Some evidence of an effect for task type was observed, although it was not entirely consistent or always very strong (see Appendix F, Table F13, for means and *SD*s for each task and task type). Two measures, K2 and off-list, favored integrated tasks (K2: F [1, 190] = 41.91, p = 0.001, *eta* = 0.18; off-list: F [1, 190] = 16.09, p = 0.001, *eta* = 0.08), while one measure, K1, favored independent tasks (F [1, 190] = 49.36, p = 0.001, *eta* = 0.21). The dependent variable for each ANOVA analysis was percentage of K1, K2, AWL, and off-list words. Effect sizes were marginal or small (see Appendix G, Table G5, for detailed results).



Note. K1 = the most frequent 1,000 words of English; K2 = the second most frequent 1,000 words of English; AWL = words from the Academic Word List; Off-list = all remaining words.

Figure 22. **Vocabulary use (2) by task.**

Phonology

Pronunciation. As noted earlier, the data for the phonology analysis came from a single independent task (Task 2) and a single integrated task (Task 3). Because Task 3 was particularly difficult, the results need to be interpreted with caution. The analysis was carried out at both word-level and subword-level, using frequency data (per 10 words or 10 syllables, respectively).

Few noticeable differences were observed in word-level pronunciation between Task 2 (independent) and Task 3 (integrated) performances (see Appendix F, Table F14, for descriptive statistics). The slightly higher mean for the independent task for the frequency of nonmeaningful words (shown in Figure 23) is perhaps attributable to the fact that speakers found it more difficult to pronounce those words in the lecture that were relatively unfamiliar (e.g., "subsidence," "San Joaquin Valley").

The result at the word-level was confirmed by the result at the subword-level (see Figure 24), where speakers produced more nontarget-like syllables in meaningful words in Task 3 (integrated) than in Task 2 (independent). A five-by-two ANOVA with the number of meaningful words on target per 10 words as the dependent variable showed this difference to be significant (F [1, 69] = 4.72, p = 0.03), although the effect size was very marginal (*eta* = 0.06; see Appendix G,

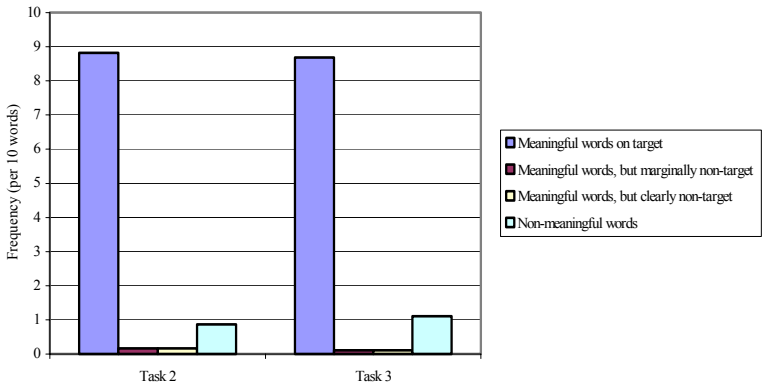


Figure 23. **Word pronunciation by task.**

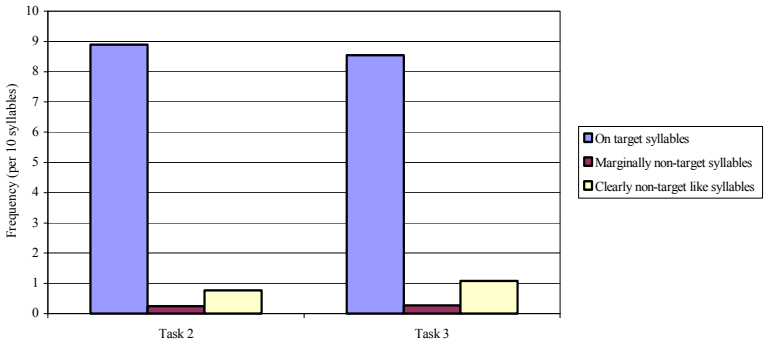
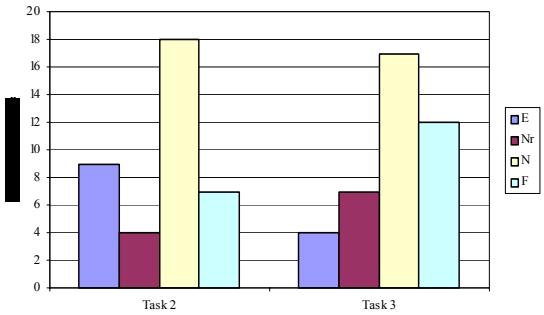


Figure 24. **Syllable pronunciation by task.**

Table G6, for ANOVA statistics). No analyses were performed on other categories as frequencies were very low.

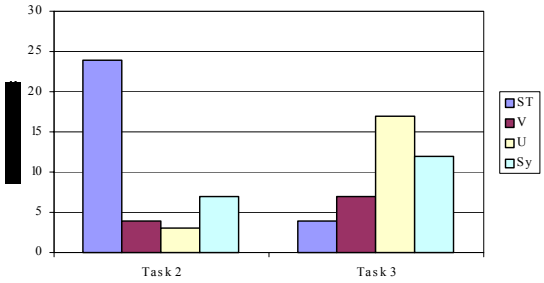
Intonation. Figure 25 presents the results for the analysis of intonation units, which indicate a mild task effect. More speakers were categorized as many and English-like for Task 2 (independent) than for Task 3 (integrated). Correspondingly, more speakers were classified as few for the integrated task. No statistical analysis was performed for this category as the observed frequencies were low (see Appendix F, Table F15).



Note. E = many and English-like; Nr = many and near English-like; N = many and non-English-like; F = few.

Figure 25. **Intonation measures by task.**

Rhythm. As Figure 26 shows, the quality of the rhythm of test-takers' speech was strikingly different in their responses to the two tasks. While much of the speech in Task 2 (independent) performances was assessed as stress-timed, approximately three-quarters of the speech on Task 3 (integrated) was classified as unclear or syllable timed. The observed frequencies precluded further statistical analysis (see Appendix F, Table F16).



Note. St = stress timed, V = variable, U = unclear, Sy = syllable timed.

Figure 26. **Rhythm measures by task.**

Fluency. Figure 27 displays the results of the analyses based on the six aspects of fluency (see Appendix F, Table F17, for means and *SD*s for each task type and task). Two of the six measures (speech rate and mean length of run) produced differences in means, both favoring independent tasks. Five-by-two ANOVA analyses with the number of filled pauses per 60 seconds, the number of unfilled pauses per 60 seconds, total pause time, the number of repairs

per 60 seconds, speech rate, and mean length of run as dependent variables confirmed that these differences were significant (speech rate: F [1, 189] = 17.27, p = 0.001, *eta* = 0.08; mean length of run: F [1, 188] = 54.54, p = 0.001, *eta* = 0.23), again with marginal or small effect sizes (see Appendix G, Table G7 for detailed results). No significant differences were observed for filled pauses, unfilled pauses, total pause time, or repairs.

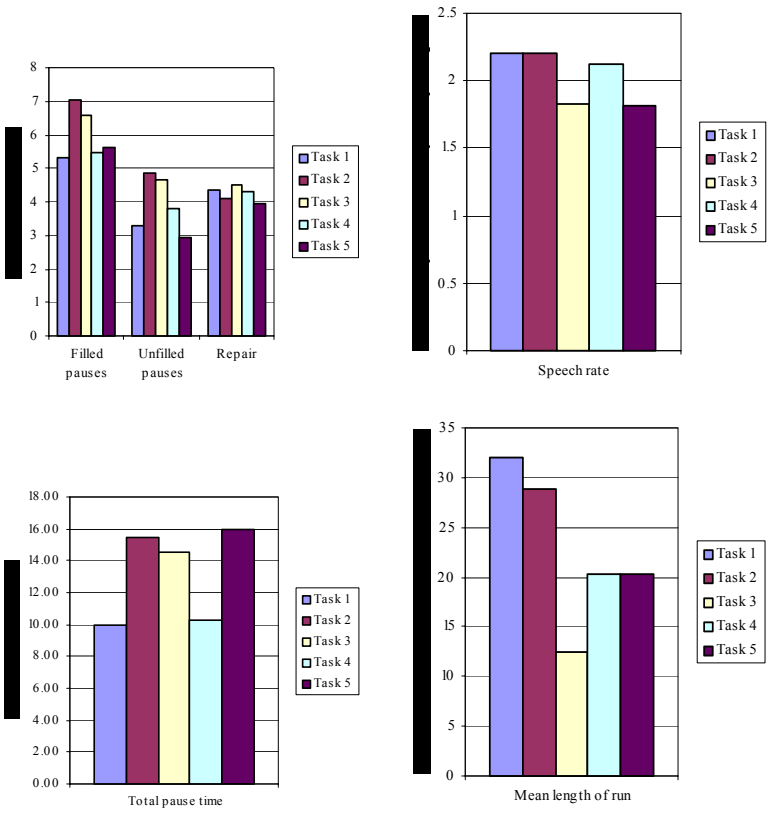


Figure 27. **Fluency measures by task.**

Content. Figure 28 presents the results for quantity of content, which was based on the number of T-units and clauses test-takers produced. ANCOVAs were run with the number of T-units and clauses per 10 utterances as dependent variables and the number of utterances as the covariate. When measured by the larger unit (i.e., in terms of T-units), we found more ideas produced for the integrated task than for the independent task (F [1, 187] = 20.68, p = 0.001, *eta* = 0.10). The reverse was found when we measured the number of ideas using the smaller unit (i.e., clauses); that is, more ideas were produced for the independent task than for the integrated task (F [1, 187] = 13.34, p = 0.01, *eta* = 0.07). This result was entirely due to the impact of the very much

smaller number of clauses in Task 3 (integrated; see Appendix F, Table F18, for means and *SD*s for each task and task type). ANCOVA statistics were used to eliminate the effect of the amount of speech (see Appendix G, Table G8, for detailed results).

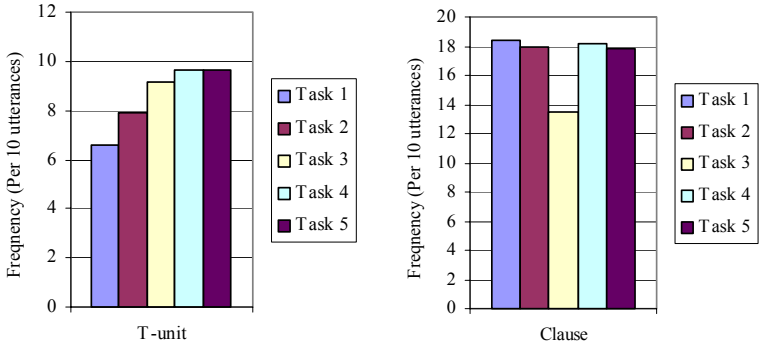


Figure 28. **Quantity of discourse by task.**

As noted earlier (and depicted in Figures 1 and 2), the schematic structure of Task 2 (independent) is simpler than that of Task 3 (integrated). Judges made comments in Study I that pointed to the differing demands of these tasks, and these were born out in the analysis of the quality of the discourse. In terms of completeness (or fulfillment of the demands of the task as described in the illustrations of their schematic structures), performances on the integrated task were more complex than performances on the independent task, but this was of course a reflection of the relative complexity of its schematic structure. In this sense, the intention of the task designers was fulfilled: The more complex tasks elicited more complex performances, as had been hoped, and this was reflected in the comments of the judges.

The schematic structure of the responses of lower-proficiency test-takers were similarly incomplete on each task, due to the limited duration of their spoken responses. The schematic structures in the performances of higher-proficiency test-takers were relatively complete on both tasks. (See Appendix I, Examples I1-I4, which presents the performances of one Level 4 and one Level 5 test-taker on Task 2 and Task 3, for an illustration. These two test-takers were able to produce speech with the expected schematic structure quite competently in both tasks.)

Summary of Results for Research Question 5

Table 17 summarizes the results of statistical analyses of performance according to task type.

Table 17

Summary of Statistical Analyses by Task Type

		Task type		Level by task	
Category	Analysis	Difference	Effect size	Difference	Effect size
Linguistic resources: Grammatical accuracy					
	Articles	√ IND>INT	0.03		
	Tense-marking				
	Third-person-singular verbs				
	Plurals				
	Prepositions				
	Global errors				
Linguistic resources: Grammatical complexity					
	T-unit complexity				
	Dependent-clause ratio				
	Verb-phrase ratio	√ IND < INT	0.02		
	Mean length of utterance				
	Modals	√ IND > INT	0.30		
	Comparatives				
	Passive				
Linguistic resources: Vocabulary					
	Word-token	√ IND > INT	0.12		
	Word-type	√ IND > INT	0.39		
	Type-token ratio	√ IND > INT	0.02		
	K1 (%)	√ IND > INT	0.21	√	0.07
	K2 (%)	√ IND < INT	0.18	√	0.06
	AWL (%)				
	Off-list (%)	√ IND < INT	0.08		
Phonology[a]					
	Meaningful words				

(Table continues)

Table 17 (continued)

Category	Analysis	Task type		Level by task	
		Difference	Effect size	Difference	Effect size
	Target-like syllables	√ IND > INT	0.06		
Fluency					
	Number of filled pauses				
	Number of unfilled pauses				
	Total pause time				
	Repairs				
	Speech rate	√ IND > INT	0.08		
	Mean length of run	√ IND > INT	0.23		
Content [a]					
	T-units	√ IND < INT	0.10		
	Clauses	√ IND > INT	0.07		

Note. √ = statistical difference; IND = Independent task; INT = Integrated task; Effect size (*eta*): marginal = (< 0.2); small = (> 0.2 < 0.5); medium = (> 0.5 < 0.8); large = (> 0.8). No statistical analyses were performed for use of logical connectives (textualization, linguistic resources category), intonation (phonology category), or rhythm (phonology category). K1 = the most frequent 1,000 words of English; K2 = the second most frequent 1,000 words of English; AWL = words from the Academic Word List; Off-list = all remaining words.

[a] Only one instance of each type of task was involved in the analysis of some or all of the aspects of this feature.

Given the different demands of the independent tasks (Tasks 1 and 2) and the integrated tasks (Tasks 3- 5), performance differences on a number of measures were investigated. The most obvious difference was in quality of ideas, where the analysis confirmed that test-takers' performances did indeed reflect the differing requirements of the tasks in terms of content and cognitive organization. In this respect, the intention of test designers in including such tasks in the test appears to have been fulfilled.

On other, more detailed measures, results were less conclusive. As stated at the beginning of this section, expectations about the impact of task type on specific discourse features were

somewhat unclear. On the one hand, the presence of input texts in the integrated tasks, and the fact that they require test-takers to communicate more cognitively complex material, might have been expected to show up in vocabulary measures and possibly also in measures of sentence complexity. On the other hand, this greater cognitive challenge might have left speakers with fewer cognitive resources available for managing grammatical accuracy, pronunciation, fluency, and rhythm, although as pointed out earlier, the findings of the literature are somewhat inconclusive on this point.

Overall, few strong differences were observed between task types; only four significant differences in which effect sizes were other than marginal were observed. One of these, mean length of run, indicated relative disfluency on the integrated task, a possibility foreshadowed in Study I. Another measure, word-type (per 60 seconds), may indicate the same.[15] This would seem to support the expectation that the greater cognitive load of the more complex, integrated tasks would lead to worse performance in terms of fluency, at least. Another measure, the percentage of vocabulary from the most frequent 1,000 words in English, indicated that a higher proportion of lexical items fell into the lowest vocabulary level in independent tasks. The results for the percentage of vocabulary from the second most frequent 1,000 words in English) and off-list vocabulary (words not on lists used in this study), although significant with a smaller effect size, confirm this, in that they indicate a lower proportion of words from the higher levels were used in the independent tasks than in the integrated tasks. This finding was again expected, because the input text in the integrated tasks provides students with task-relevant vocabulary.

In contrast with these expected findings, one of the grammatical complexity measures, modals (an aspect of sophistication), favored independent tasks rather than integrated ones. The greater use of modals in the independent tasks can be explained by the different types of targeted functions that were required for the different task types. In the independent tasks, the functions of speech were opinion for Task 1 and value/significance for Task 2 (as described earlier in Table 1). Modals were a frequent feature of the speech samples (see Appendix J for examples). On the other hand the functions in the three integrated tasks were explain/describe/recount. To be able to explain/describe/recount, the test-takers did not necessarily need to use modals to express their opinions as frequently as they did in the independent tasks. As the examples in Appendix J show, very few modals were observed in the integrated task performances. The absence of a global effect for grammar confirms the findings of Study I, in which differences in comments on grammar were

not generalized across task types, but were restricted to grammatical constructions that were specific to particular texts.

Task 3 appeared to be a particularly demanding task. This was reflected in lower measures for Task 3 on a number of the performance features and in the scores test-takers achieved. A comparison of the scores awarded to test-takers who completed and received scores for all five tasks (N = 274) revealed that average scores were almost half a score-level[16] lower on Task 3 than on the other tasks. Furthermore, as shown in Table G5 (Appendix G), we found significant interaction effects for level and task type on two measures (i.e., K1 and K2). The impact of Task 3 on the analysis is heightened by the fact that for the phonology measures (pronunciation, intonation, and rhythm), Task 3 was the only integrated task analyzed.

The findings of this study are relevant to the currently much-debated issue of the impact of cognitively demanding tasks on features of performance. As we have seen, the findings of research on the information-processing approach to tasks proposed by Skehan (1996, 1998) initially led us to expect a possible difference in the quality of performance between integrated and independent tasks. Overall, however, we did not find performance differences between task types on most measures of the features most frequently studied in the work in this tradition—that is, grammatical accuracy, grammatical complexity, and fluency (with the exception of one of the fluency measures). Interestingly, this reflected findings in earlier research on the impact of task demands on test-taker performance (McNamara et al., 1999, Iwashita et al., 2001), which reported similar absence of effects for performance conditions that differed in their cognitive demands.

Overall, the analysis by task type showed an obvious impact on the type of content, as predicted, but this was not matched by other content quality measures, particularly vocabulary. Some specific differences in linguistic processing were observed for the different task types, with some modest evidence of an impact on fluency, but otherwise strong differences were not observed, which is interesting in light of the ongoing debate about the impact of cognitive load on features of performance in speaking tasks. The Study II analysis corresponded on a number of detailed points with the findings of the Study I, as indicated above. From a practical point of view, the validity of the integrated tasks appears to be mostly a question of task content. In task design, the functional demands and the difficulty of specific tasks need to be carefully anticipated and monitored in performance.

Interpretation and Discussion of Findings

It has been argued that scoring descriptors should closely match what raters perceive in the performances they have to grade, and that "the starting point for scale development should surely therefore be a study of the perceptions of proficiency by raters in the act of judging proficiency" (Pollitt & Murray, 1996, p. 76). While there is a growing body of research involving verbal reports produced by assessors of writing, and these verbal reports have long been argued to provide a sound basis for rating-scale development, research involving speaking tests is relatively scarce. In Study I, which examined verbal reports produced by field experts to inform the development or validation of rating scales for the assessment of EAP speaking, we found that all judges focused on the same general categories and tended to discuss the components of these categories in essentially similar ways. This indicates a broad level of agreement on the construct, at least as it pertains to performance on the tasks used here.

We also found that there was a greater focus on the content of test-takers' responses than had been found in other verbal-report studies involving speaking tasks (Brown, 2000; Meiron, 1998; Pollitt & Murray, 1996), perhaps because these were all concerned with nonacademic monologic or interview tasks. In this respect, the assessment of the academic speaking tasks in the current study resembles the assessment of academic writing, in which the quality of the content is typically specified as one of the main criteria as well as a major focus of raters (see, for example, Cumming et al., 2002). Somewhat surprisingly, however, in terms of the percentage of comments devoted to the content of test-takers' speech samples, the focus on content was as great in the independent tasks as it was in the integrated ones.

Other than content, aspects of performance that were highly salient to judges included the more traditional linguistic resources (i.e., grammar and vocabulary), along with production features, such as fluency and pronunciation. Linguistic resources were typically assessed in terms of sophistication or complexity on the one hand, and accuracy on the other. Hesitation and repair contributed to assessments of fluency, and intonation and rhythm were assessed in addition to pronunciation. Particularly marked was the emphasis placed on comprehensibility or clarity as opposed to correctness or nativeness, which occurred in all aspects of performance—phonological, syntactic, and organizational. The richness of the verbal report data—together with the fact that the judges tended to heed the same categories, display similar ways of talking about performances using similar terminology,[17] and produce descriptions that captured differences between increasing

levels of proficiency—points to the usefulness of this type of data for the construction of "assessor-friendly" rating scales (Alderson, 1991).

In order to examine the validity of the draft scales used to provide the baseline score data for the performances used in this study, the categories derived from this analysis were compared with those reflected in the draft scales. Table 18 lists features that were singled out for attention in the scales and confirmed in the verbal-report analysis. The evidence suggests that the verbal reports collected here on the whole validate the ETS scales in terms of the salience of the features nominated within them to field experts. In addition, one of the most noticeable features of the ETS scales is the emphasis on the comprehensibility or clarity of test-takers' speech (i.e., the impact on the listener), rather than on accuracy or nativeness. This feature echoes the findings of the verbal-report analysis. The usefulness of the verbal-report data in the validation of the draft ETS scales in the present study indicates the appropriateness of the approach more generally in test validation studies.

Given the focus in this study on integrated tasks as a novel approach to the assessment of oral ability, another question concerned the need for task- or task-type-specific criteria. In fact we found that, in general, all judges attended to all of the conceptual categories across all tasks and task types, which points towards the utility of a general scoring framework that can be applied to all tasks. However, the findings of Study I indicated that the content category was the most task- and task-type specific; they looked for specific functions, specific content, organization, and text structures, which differed by task type, but also by task to some extent. These differences were supported by the findings of Study II: Performance on the different task types differed the most in terms of the functional skills and rhetorical structure of the spoken texts, and the content was text-specific.

Table 18

Mapping of Judges' Conceptual Categories and ETS Scale Descriptions

ETS draft scales	Verbal report categories
Independent tasks	*Independent tasks*
Clear viewpoint, reasons, examples	Opinion, reasons, examples[a]
Relevance/sophistication of ideas ("well-chosen")	Relevance/sophistication of ideas[a] (content)
Organization of thoughts	Logic/clarity of argument[a] (content)

(Table continues)

Table 18 (continued)

ETS draft scales	Verbal report categories
Integrated tasks	*Integrated tasks*
Major ideas, important supporting ideas	Main ideas, supporting detail, relevance[a]
Accuracy of information	Accuracy[a]
Organization of thoughts	Organization of ideas[a]
Fluency/hesitation	Fluency/pauses[b]
Intelligibility/listener effort	Intelligibility/clarity[c]
Vocabulary and grammar—range and accuracy	Vocabulary and grammar—range and accuracy[d]
Completeness of thoughts	Completeness of utterances[d]
Verbatim use of source text	Dependence on input text/prompt[d]

[a] Content. [b] Fluency. [c] Pronunciation. linguistic resources. [d] Linguistic resources.

While this finding provides evidence for the value of including the different task types in assessments of oral proficiency, the different task structures also indicate again that task- or task-type-specific scales appear to be warranted for the content dimension of the performance. Moreover, as the evidence here supports Cohen's (1993) finding that raters did not fully agree on which ideas were essential to the construction of a meaningful summary, we argue that there is a strong need for guidance and training in assessing on-task performance, particularly in relation to content quality. At least some level of additional task-specific detail is needed to accompany the general descriptors in order to minimize disagreement as to content appropriateness. This recommendation echoes that of Cumming et al. with respect to integrated writing tasks (2001, p. 70). This level of detail could take the form of an outline of expected schematic structure or annotated sample responses that explain the relationship between content and proficiency level. Given that different schematic structures are expected for different tasks, and that (for integrated tasks at least) these are specific to individual input texts, we recommend that guidance be given in relation to each individual task prompt.

A further point of difference in independent and integrated tasks concerns the role of comprehension in productive performance on integrated tasks. We noted earlier that the judges in this study, who were working without the guidance of specific criteria or scales, appeared uncertain as to whether to compensate for the heavy cognitive load that integrated tasks imposed (in terms of understanding, remembering, and transforming the input text). They made frequent reference to the effect of lack of comprehension on the quality of the information contained in the responses (e.g., a

lack of detail or accuracy, or poor organization of ideas and therefore lack of coherence).

Disfluency was also attributed to comprehension or recall problems. This suggests that rather than offering the opportunity to tap test-takers' ability to produce complex speech, the processing requirements appear to impact on their ability to produce fluent speech and well-structured content. Judges' comments indicated specifically that they felt the task interfered with test-takers' ability to display their "real" level of proficiency. (A comparison of scores awarded on different tasks to the same test-takers indicated that they scored lower, in general, on Task 3, the integrated listening-speaking task, than on the others.) While this might raise questions about the validity of integrated tasks as measures of speaking proficiency, in practical terms it points again to a need for clear guidance and training for raters in order to minimize potential disagreement. We do also note that this problem was far less marked in the data used in the current study than in the data used in an earlier study (Brown et al., 2001); in the earlier case, the input texts for the integrated tasks were considerably more difficult and problematic. This would suggest that in order to ensure that test performance is not unduly affected by the difficulty of particular input texts, they should be carefully vetted for comprehensibility and equivalence.

Study II sought evidence of the empirical basis of raters' perceptions through an analysis of test-taker discourse. In general, the findings of Study II supported the findings of Study I, providing validity evidence for the rating categories that were identified and, because of the degree of overlap, for the draft scales developed by ETS. In reconciling the findings of the discourse analysis with the verbal report categories, we found that there were trends in the expected direction for all of the major conceptual categories, with, generally, significant differences across the levels for individual features. However, trends for some features were more marked than others, and some features did not differ significantly across levels. Where they did differ, the standard deviations for most measures tended to be large and the effect sizes for interlevel significance tests were generally small, indicating that individual performances at any one level for any one feature of analysis varied considerably.

This lack of clear distinction between performances at adjoining levels is perhaps attributable to the use of a holistic assessment to provide the baseline score data, rather than more specifically focused analytic scores. It is possible that, if the significance tests were to be re-run using scores derived from analytic ratings, more conclusive evidence of the validity (discreteness) of the levels would emerge. Moreover, given the number of individual analyses all contributing to

the measurement of the same main conceptual category, it is hardly surprising that differences between levels on any single measure are not very marked. A further point that might also help explain the lack of discreteness of the levels, in terms of the analyses undertaken in this study, is the previously discussed tension that exists between measurable features (such as accuracy or complexity) and essentially subjective and unmeasurable ones, such as intelligibility or clarity (that is, those that relate to the impact of problems in the performance on the reader). It may be that the specific measures used in this study, which necessarily focus on quantifiable features, do not reflect the criteria most salient to judges.

Given the interest in this study in the functioning of integrated tasks, the question was posed as to whether the different task types would result in performances that differed in quality in ways other than the type of content. Two apparently opposing sets of expectations were presented: Would the greater cognitive demand posed by the integrated tasks result in responses, at least at the higher levels, that were more sophisticated and complex (Robinson 1995; 1996; 2002)? Or would the challenge prove to be a source of difficulty that leads speaker to produce *less* sophisticated, complex, accurate, and possibly fluent responses (Skehan, 1996, 1998)?

In fact both of these expectations were partially upheld, although it was not the case that the findings were the same for all tasks within a task type. There was some evidence that the level of vocabulary was slightly higher in the integrated tasks, but little evidence of differences in structural sophistication or complexity. Whereas it had been expected that the cognitive complexity of the task (comprehending, recalling, and organizing input material) would have an adverse impact on the quality of integrated-task performance, the speech samples analysis revealed that this was the case only for Task 3, and mainly for fluency. This, we speculate, was a text-induced difference, derived from (a) the complexity of ideas and (b) the linguistic density of the input text, which combined with the one-time hearing and concomitant recall demands of a listening text, presented the greatest challenge of the three integrated tasks to test-takers. This finding may indicate a lower text complexity threshold for listening input texts than for reading input texts. In order to ensure the validity of the task as a measure of speaking skill, it would seem necessary to ensure that the level of input text difficulty falls below a threshold that enables all but the very weakest students to comprehend and recall relatively readily. It also points to the need to monitor relative input text difficulty across test versions.

In terms of our original assumptions, then, there appears to be little evidence to support the

view that task performance will be more linguistically complex or sophisticated on integrated tasks than on independent tasks. However, if the aim of integrated tasks is essentially to provide a context for the demonstration of more complex text structuring, which was found to occur, this may not be an issue.

What evidence there is of richer vocabulary is limited and is, of course, likely to reflect the re-use of input vocabulary rather than the "original" use of sophisticated lexis; this emerged as an assessment problem for raters in that inconsistency was noted as to which should be valued: evidence that candidates are able to re-use input vocabulary appropriately or whether they are able to provide alternatives. Because these orientations could potentially lead to different ratings outcomes, they raise concerns about both validity (what is the vocabulary construct being assessed?) and fairness (will candidates be disadvantaged by encountering one rater rather than another?). In relation to validity, the strategy of re-using vocabulary (or indeed, longer units of input text) may be taken as evidence of a certain type of linguistic or strategic communication skill; it may also be conceptualized negatively as evidence of a lack of alternative vocabulary, or as plagiarism (although as Lumley, Brown, & Zhang, 2002, showed in relation to integrated writing tasks, the question of plagiarism, particularly in relatively narrowly constrained and inauthentic test tasks such as these, is far from straightforward). Whatever the case, both raters and learners need to be made aware of the expectations.

In conclusion, this study has shown the usefulness of two types of analysis—verbal report analysis and discourse analysis—in the development and/or validation of speaking tasks. In particular, this study showed that domain experts were able to distinguish and describe qualitatively different performances using a common set of criteria very similar to those included in draft scales developed for the tasks at ETS. Analysis of test-taker discourse in relation to measures selected to reflect criteria applied by the judges revealed empirical support for the categories. Limitations of this study pertained to the strength of the differences for specific measures—a function, as argued earlier, of the fact that performances were assessed only holistically. While the limited usefulness of some commonly used linguistic measures when analyzing spoken data and analyzing short segments of speech has been commented on earlier, here we reiterate the value of undertaking what are essentially very complex and labor-intensive analyses in the validation of high-stakes tests such as TOEFL.

Implications for TOEFL

The two coordinated studies reported in this document have important implications for the use of integrated tasks in the context of the new TOEFL examination. These pertain to the usefulness and design of integrated tasks, to the development of assessment scales, and to rater training and candidate preparation. While these have been alluded to above, here the main implications are set out as a series of recommendations:

1. Both the use of verbal report data produced by field experts and analyses of actual task performance can contribute to the development of valid, user-friendly scales.
2. The study demonstrates the utility of integrated tasks in allowing candidates to demonstrate more complex functional and text organizational skills than is the case with independent tasks.
3. The study confirms the relevance of the criteria contained within the draft ETS scales for the assessment of proficiency on integrated tasks. It should be noted that the assessment of content quality is particularly relevant to judgments of EAP proficiency in both integrated and independent and tasks.
4. Task-type-specific scales appear to be warranted for content. Moreover, because the text structure and selection of ideas is text- rather than task-type specific, it is recommended that raters be given additional guidance on content assessment through the use of sample responses and/or schematic structure models.
5. It is important to monitor the difficulty of input texts carefully in order to ensure that the time-constrained processing demands do not adversely affect the quality of speech. Text difficulty equivalence for alternative task versions will need to be carefully monitored through piloting in order to ensure fairness. Because of these potential problems, it is also recommended that integrated tasks are not used exclusively for the assessment of EAP speaking proficiency.
6. The integration of comprehension and production in integrated tasks complicates the rater's task. Where holistic scales are used to assess such performance it would appear necessary to provide guidance to raters on how to manage their perceptions of the interdependency of these skills.

References

Alderson, J. C. (1991). Bands and scores. In J. C. Alderson & B. North (Eds.), *Language testing in the 1990s: The communicative legacy* (pp.71-86). London: Macmillan.

Bachman, L. F., & Savignon, S. J. (1986). The evaluation of communicative language proficiency: A critique of the ACTFL oral interview. *Modern Language Journal, 70*, 380-390.

Berman, R. A., & Slobin, D. I. (1994). Filtering and packaging in narrative. In R. A. Berman & D. I. Slobin (Eds.), *Relating events in narrative: A crosslinguistic developmental study* (pp. 515-554). Mahwah, NJ: Lawrence Erlbaum Associates.

Brindley, G. (1986). *The assessment of second language proficiency: Issues and approaches*. Adelaide, Australia: National Curriculum Resource Centre.

Brindley, G. (1991). Defining language ability: The criteria for criteria. In S. Anivan (Ed.), *Current developments in language testing* (pp. 139-164). Singapore: SEAMEO Regional Language Centre.

Brown, A. (2000). An investigation of the rating process in the IELTS Speaking Module. In R. Tulloh (Ed.), *Research reports* (1999, Vol. 3, pp. 49-85). Canberra, Australia: IELTS Australia.

Brown, A. (2003). Interviewer variation and the co-construction of speaking proficiency. *Language Testing, 20*(1), 1-25.

Brown, A., Elder, C., Lumley, T., McNamara, T., & McQueen, J. (1992). Mapping abilities and skill levels using Rasch techniques. *Melbourne Papers in Language Testing 2*(2), 35-54.

Brown, A., & Hill, K. (1998). Interviewer style and candidate performance in the IELTS oral interview. In S. Woods (Ed.), *Research reports* (1997, Vol. 1, pp. 1-19). Sydney, Australia: ELICOS Association and IELTS Australia.

Brown, A., McNamara, T., Iwashita, N., & O'Hagan, S. (2001). *Investigating raters' orientations in specific-purpose, task-based oral assessment*. Unpublished manuscript.

Brown, R. (1973). *A first language: The early stages.* Cambridge, MA: Harvard.

Cafarella, C. (1997). Assessor accommodation in the VCE Italian Oral Test. *Australian Review of Applied Linguistics, 20*(1), 21-41.

Cobb, T. (2002). *The Web Vocabulary Profiler* (Version 1.0) [Computer program]. University of Québec, Montréal. Retrieved January 12, 2005, from http://www.er.uqam.ca/nobel/r21270/textools/web_vp.html

Cohen, A. (1993). The role of instructions in testing summarizing ability. In D. Douglas & C. Chapelle (Eds.), *A new decade of language testing research: Selected papers from the 1990 Language Testing Research Colloquium* (pp.132-160). Alexandria, VA: TESOL Publications.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34*(2), 213-238.

Cumming, A., Kantor, R., & Powers, D. E. (2001). *Scoring TOEFL essays and TOEFL 2000 prototype writing tasks: An investigation into raters' decision-making and development of a preliminary analytic framework* (TOEFL Monograph No. MS-22). Princeton, NJ: ETS.

Cumming, A., Kantor, R., & Powers, D. E. (2002). Decision-making while rating ESL/EFL writing tasks: A descriptive framework. *The Modern Language Journal, 86*, 67-96.

Cumming, A., Kantor, R., Powers, D., Santos, T., & Taylor, C. (2000). *TOEFL 2000 writing framework: A working paper* (TOEFL Monograph No. MS-18). Princeton, NJ: ETS.

Douglas, D. (1994). Quantity and quality in speaking test performance. *Language Testing, 11*(2), 125-144.

Douglas, D., & Selinker, L. (1992). Analyzing oral proficiency test performance in general and specific-purpose contexts. *System, 20*, 317-328.

Douglas, D., & Selinker, L. (1993). Performance on a general versus a field-specific test of speaking proficiency by international teaching assistants. In D. Douglas & C. Chapelle (Eds), *A new decade of language testing research* (pp. 235-256). Alexandria, VA: TESOL Publications.

Eggins, S., & Slade, D. (1997). *Analyzing casual conversation.* London: Cassell.

Enright, M., Bridgeman, B., & Cline, F. (2002, April). *Prototyping a test design for new TOEFL*. Paper presented at the meeting of the National Council on Measurement in Education, New Orleans, LA.

Ericsson, K., & Simon, H. (1993). *Protocol analysis: Verbal reports as data* (Rev. ed.). Cambridge, MA: MIT Press.

Foster, P., & Skehan, P. (1996). The influence of planning and task type on second language performance. *Studies in Second Language Acquisition, 18*(3), 299-323.

Freed, B. (2000). Is fluency, like beauty, in the eyes (and ears) of the beholder? In H. Riggenbach (Ed.), *Perspectives on fluency* (pp 243-265). Ann Arbor: University of Michigan Press.

Fulcher, G. (1987) Tests of oral performance: The need for data-based criteria. *ELT Journal 41*, 287-291.

Fulcher, G. (1993). *The construction and validation of rating scales for oral tests in English as a foreign language*. Unpublished doctoral dissertation, University of Lancaster, England.

Fulcher, G. (1996). Does thick description lead to smart tests? A data-based approach to rating scale construction. *Language Testing, 13*(2), 208-238.

Green, A. (1998). *Verbal protocol analysis in language testing research: A handbook* (Studies in language testing, Vol. 5). Cambridge, England: Cambridge University Press.

Griffin, P. (1990). *Profiling literacy development: Monitoring the accumulation of reading skills.* Melbourne, Australia: Assessment Research Centre, Phillip Institute of Technology.

Hamilton, J., Lopes, M., McNamara, T. F., & Sheridan, E. (1993), Rating scales and native speaker performance on a communicatively-oriented EAP test. *Language Testing, 10*(3), 337-353.

Hunt, K. W. (1970). Syntactic maturity in school children and adults. *Monographs of the Society for Research in Child Development, 35*(1). (1, Serial No. 134)

Huot, B. (1993). The influence of holistic scoring procedures on reading and rating student essays. In M. Williamson & B. Huot (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations* (pp. 206-236). Cresskill, NJ: Hampton Press.

Iwashita, N. (1996). The validity of the paired interview in oral performance assessment. *Melbourne Papers in Language Testing, 5*(2), 51-65.

Iwashita, N., McNamara, T., & Elder, C. (2001). Can we predict task difficulty in an oral proficiency test? Exploring the potential of an information processing approach to task design. *Language Learning, 21*(3), 401-436.

Johnson, M. (2000). Interaction in the Oral Proficiency Interview: Problems of validity. *Pragmatics, 10*(2), 215-231.

Lantolf, J., & Frawley, W. (1988). Proficiency: Understanding the construct. *Studies in Second Language Acquisition, 10*(2), 181-96.

Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics, 16*(3), 307-322.

Lazaraton, A. (1996). Interlocutor support in oral proficiency interviews: The case of CASE. *Language Testing, 13*(2), 151-171.

Lewkowicz, J. (1997). The integrated testing of a second language. In C. Clapham & D. Corson (Eds.), *Encyclopaedia of Language and Education* (Vol. 7: Language Testing and Assessment, pp.121-130). Dortrecht, The Netherlands: Kluwer.

Lumley, T. (2000). *The process of the assessment of writing performance: The rater's perspective.* Unpublished doctoral dissertation, University of Melbourne, Australia.

Lumley, T., Brown, A., & Zhang, W. X. (2002, December). *Learner responses to new TOEFL integrated reading/writing tasks.* Paper presented at 25th annual Language Testing Research Colloquium, Hong Kong, China.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk* (Vol. I: Transcription format and programs, 3rd ed.). Mahwah, NJ: Lawrence Erlbaum.

Martin, J. R. (1992). *English text: System and structure*. Amsterdam: John Benjamins.

Martin, J. R., & Rothery, J. (1986). What a functional approach to the writing task can show about "good writing.' In B. Couture (Ed.), *Functional Approaches to Writing* (pp. 241-265). Norwood, NJ: Ablex.

Matthews, M. (1990). The measurement of productive skills: Doubts concerning the assessment criteria of certain public examinations. *ELT Journal, 44*(2), 117-121.

McNamara, T. (1990). Item response theory and the validation of an ESP test for health professionals. *Language Testing, 7*(1), 52-75.

McNamara, T. F. (1996). *Measuring second language performance*. London and New York: Longman.

McNamara, T., Elder, C., & Iwashita, N. (1999). *Investigating predictors of task difficulty in the measurement of speaking proficiency* (Final Report, TOEFL 2000 Research Project). Unpublished report, University of Melbourne, Australia.

McNamara, T., Hill, K., & May, L. (2002). Discourse and assessment. *Annual Review of Applied Linguistics, 22*, 221-242.

Mehnert, U. (1998). The effects of different lengths of time for planning on second language performance. *Studies in Second Language Acquisition, 20*(1), 83-106.

Meiron, B. E. (1998). Rating oral proficiency tests: A triangulated study of rater thought processes. Unpublished master's thesis, University of California at Los Angeles.

Milanovic, M., & Saville, N. (1994, March). *An investigation of marker strategies using verbal protocols*. Paper presented at the 16th annual conference of the Language Testing Research Colloquium, Washington, DC.

Milanovic, M., Saville, N., & Shen, S. (1996). A study of the decision-making behavior of composition markers. In M. Milanovic & N. Saville (Eds.), *Performance testing, cognition and assessment: Selected papers from the 15th Language Testing Research Colloquium, Studies in Language Testing 3* (pp. 92-114). Cambridge, England: Cambridge University Press.

Mullis, I. V. S. (1980 & 1981). *Using the primary trait system for evaluating writing*. (ETS Research Report No. 10-W-51). Princeton, NJ: ETS.

Norris, J. (2001). Identifying rating criteria for task-based EAP assessment. In T Hudson & J. D. Brown (Eds.), *A focus on language test development: Expanding the language proficiency construct across a variety of tests* (Technical Report No. 21, pp.163-204). Honolulu: University of Hawaii, Second Language Teaching and Curriculum Center.

North, B., & Schneider, G. (1998). Scaling descriptors for language proficiency scales. *Language Testing, 15*(2), 217-263.

O'Loughlin, K. (1997). The comparability of direct and indirect speaking tests: A case study. Unpublished doctoral dissertation, University of Melbourne, Australia.

O'Loughlin, K. (2002). The impact of gender in oral proficiency testing. *Language Testing, 19*(2), 169-192.

Ortega, L. (1998). Using CHILDES for the analysis of L2 speech: Task variation, syntactic complexity, and lexical production. Manoa: University of Hawaii at Manoa.

Ortega, L. (1999). Planning and focus on form in L2 oral performance. *Studies in Second Language Acquisition, 21*(1), 109-148.

Ortega, L., Iwashita, N., Rabie, S.. & Norris, J. (1998). Coding guidelines: T-unit and clauses segmentations. Unpublished manuscript, University of Hawai'i at Manoa.

Ortega, L., Rabie, S., Iwashita, N., & Norris, J. (1998). Transcription guidelines for cross-linguistics analysis. Unpublished manuscript, University of Hawai'i at Manoa.

Paltridge, B. (2000). *Making sense of discourse analysis*. Gold Coast, Australia: Gerd Stabler.

Pica, T. (1983). Methods of morpheme quantification: Their effect on the interpretation of second language data. *Studies in Second Language Acquisition, 6*(1), 69-79.

Pienemann, M., & Johnson, M. (1987). Factors influencing the development of language proficiency. In D. Nunan (Ed.), *Applying second language acquisition research* (pp. 89-94). Adelaide, Australia: National Curriculum Resource Centre.

Pollitt, A., & Murray, N. L. (1996). What raters really pay attention to? In M. Milanovic & N. Saville (Eds.), *Performance testing, cognition and assessment: Selected papers from the 15th Language Testing Research Colloquium* (pp. 74-91). Cambridge, England: Cambridge University Press.

Robinson, P. (1995). Task complexity and second language narrative discourse. *Language Learning, 45*(1), 99-140.

Robinson, P. (1996). Introduction: Connecting tasks, cognition and syllabus design. *The University of Queensland Working Papers in Language and Linguistics, 1*(1), 1-15.

Robinson, P. (2002). Task complexity, cognitive resources, and second language syllabus design. In P. Robinson (Ed.), *Cognition and second language instruction* (287-318). New York: Cambridge University Press.

Rommark, K. (1995). *Xwaves.* Los Angeles: University of California, Los Angeles. http://www-ssc.igpp.ucla.edu/~bryan/xwaves

Ross, S., & Berwick, R. (1992). The discourse of in oral proficiency interviews. *Studies in Second Language Acquisition*, *14*(2), 159-176.

Shohamy, E. (1994). The validity of direct versus semi-direct oral tests. *Language Testing, 11*(2) 99-123.

Skehan, P. (1996). A framework for the implementation of task-based instruction. *Applied Linguistics, 17*(1), 38-62.

Skehan, P. (1998). *A cognitive approach to language learning*. Oxford: Oxford University Press.

Skehan, P., & Foster, P. (1999). The influence of task structure and processing conditions on narrative retellings. *Language Learning*, *49*(1), 93-120.

Strauss, A., & Corbin, J. (1994). Grounded theory methodology: An overview. In N. K. Denzin & Y. S. Lincon (Eds.). *Handbook of qualitative research.* London: Sage.

Towell, R., Hawkins, R., & Bazergui, N. (1996). The development of fluency in advanced learners of French. *Applied Linguistics, 17*(1), 84-119.

Upshur, J., & Turner, C. (1995). Constructing rating scales for second language tests. *ELT Journal, 49*, 3-12.

van Lier, L. (1989). Reeling, writhing, drawling, stretching, and fainting in coils: Oral proficiency interviews as conversation. *TESOL Quarterly, 23*(3), 489-508.

Vaughan, C. (1991). Holistic assessment: What goes on in the rater's mind? In L. Hamp-Lyons (Ed.), *Second language writing in academic contexts* (pp. 111-125). Norwood, NJ: Ablex.

Weigle, S. C. (1994). Effects of training on raters of ESL compositions. *Language Testing*, *11*(2), 197-223.

Wigglesworth, G. (1997). An investigation of planning time and proficiency level on oral test discourse. *Language Testing, 14*(1), 85-106.

Wolfe-Quintero, K., Inagaki, S. & Kim, H-Y. (1998). *Second language development in writing: Measures of fluency, accuracy, and complexity* (Technical Report No. 17). Honolulu: Second Language Teaching & Curriculum Center, University of Hawaii.

Young, R., & He, A. W. (1998). *Talking and testing: Discourse approaches to the assessment of oral proficiency*. Amsterdam: John Benjamins.

Notes

[1] For a more detailed discussion of Fulcher's study, see Discourse Studies in Oral Assessment later in this report.

[2] This was not always possible at the lowest level.

[3] No overall value of kappa could be calculated for these categories (or for any subpart, such as grammar), since at this level, categories such as grammar and vocabulary were not unique. That is, at this level, the (countable) string of discourse could be classified by one or more descriptions.

[4] From this point forward, italicized terms are those commonly used by judges.

[5] Although it is likely that the term "pronunciation" was frequently used as a cover term for both segmental pronunciation and suprasegmental prosodic marking.

[6] The disjunction between language and content was nevertheless less commonly as frequent or problematic as in the previous study (Brown et al., 2001), in which failure to understand the input texts was frequently judged as interfering with test-takers ability to complete the speaking task.

[7] This was naturally more the case in the reading-speaking task than the listening-speaking tasks.

[8] Word-types were classified according to both semantic and syntactic categories. For example, the sentence "I disagree with the point of views shown in the newspaper" (Level 3, Task 1) has 11 word-tokens and 10 word-types.

[9] Repair refers to repetition of exact words, syllables, or phrases; replacement; reformulations (grammatical correction of structural features); false starts; and partial repetition of some part of a word or utterance (Freed, 2000).

[10] An intonation unit is a stretch of speech produced under a single intonation contour. That is to say, they are often, though not always, preceded and followed by pauses; they may be marked by initial pitch reset. In the major varieties of English they invariably contain prosodic lengthening on the final syllable and carry distinctive phrase-final melodies; internally, a restricted set of melodic constituents also makes for a "well-formed" intonation phrase, and syllables are timed relative to one another in a particular way. The sentence, "What did you put in my |drink| Jane?" contains three intonation units ("what did you put in my," "drink," and "Jane").

[11] Pauses of 3 or more seconds are considered to be a substantial length of elapsed time, and are therefore excluded from total speaking time.

[12] We considered initially using multivariate statistics (i.e., MANOVA, MANCOVA) for each category under investigation (e.g., grammatical accuracy, grammatical complexity, fluency) instead of univariate statistics for each aspect in the category (e.g., global error and tense-marking in grammatical accuracy). However, after conducting trial analyses and consulting with a statistician, it was decided to use univariate statistics as the differences in the results of the trial univariate and multivariate analyses were not very large.

[13] The measures that served as the dependent variable were the number of clauses per T-unit (T-unit complexity), the ratio of dependent clauses to the total number of clauses (dependent-clause ratio), the number of verb phrases per T-unit (verb-phrase ratio), and the number of morphemes per utterance (mean length of utterance). The covariate was the number of utterances.

[14] Before carrying out final analyses, multivariate analyses were conducted by putting all aspects of a category (e.g., for grammatical accuracy: tense marking, article use, plural, third-person-singular, and prepositions) together, as were univariate analyses (i.e., each aspect of one category was investigated separately). However, the results yielded from univariate and multivariate analyses were not markedly different.

[15] The meaning of this measure is ambiguous. A lower measure of word-types per 60 seconds could reflect a slower rate of speaking or it could reflect less lexical diversity. However, as the type-token ratio for the two task types is only marginally significant, this does not appear to be the case here.

[16] On the scale of 0-5 used to assess performance

[17] Although the degree of consistency among raters as to their rating orientations was not a specific focus of the study, an impressionistic review of the coded data revealed substantial agreement in the types of features on which they focused.

List of Appendices

Page

Appendix A
Task Prompts

Task 1

Recently there have been proposals in some cities that high school students should attend classes twelve months a year. If this were adopted, what effects do you think it would have on these communities?

Targeted function and content: Opinion; impersonal focus; factual/conceptual information

Task 2

The newspaper headline below reflects a situation faced by an increasing number of school systems. What is your opinion about the value of art and music courses in a student's educational training?

Targeted function and content: Value/significance; impersonal focus; factual/conceptual information

Task 3

The professor describes a series of events that occurred in the San Joaquin Valley in California. Explain what happened there. In your response, be sure to include details about:

- the problems that occurred there
- the causes of the problems
- the efforts that were made to solve the problems

Targeted functions: Explain/describe/recount

Targeted discourse features: Example/event; cause/effect

Task 4

The student and tutor discuss an experiment. Describe the details of the experiment and explain what the results of the experiment show.

Targeted functions: Explain/describe/recount

Targeted discourse features: Process/procedure; Purpose/results

Task 5

Describe Robert Fantz's experiment [in which he used the visual stimuli below]. In your response, include information about the following:

- the purpose of the experiment
- important details of the experiment

Targeted functions: Explain/describe/recount

Targeted discourse features: Process/procedure; purpose/results

Appendix B
Sample Verbal Report

Summary Turn

This speaker has a clear expression of her opinion although she has fairly limited resources in terms of grammar. My impression would be that she's perhaps a long-term resident who's been using limited English for some time. Her pronunciation is not perfect but the problems are not intrusive, and she has enough use of stress intonation to communicate her opinions. Her speaking's a little bit monotonous to listen to but she is getting her opinion across, and she has organized her thoughts quite well. So despite the limitations in grammar and vocabulary, this is a reasonably developed expression of opinion. My feeling would be that it would be difficult to eliminate some of those language weaknesses.

Review Turns

Now I didn't hear any *if*, I don't know whether that was missing from the beginning of the tape—*if high school students attend the classes*—it may be that the speaker left it out.

It's going to be hard for them, "if they attend it's going to be"—she using future rather than conditional.

Because the high school student—she leaves the "s" off the end of her word.

She has a definite opinion about how many breaks, but that's not unreasonable.

So she states the situation and then introduces her opinion about the effect with good intonation. The effect sets up what she wants to say next.

Now the grammar wasn't strictly appropriate but she's managed to get into the point, the effect, cause, the community.

So they don't like, *they would effected* rather than "they would be affected," but the meaning is coming across clearly.

So she's thinking about the effect on the individual and the wider effect on the community.

And then she returns to the first point, the conclusion, so she has a conclusion—*it would be better*.

And she refers to the benefits of having the break—that they'd be *fresh in the next semester*. So she has some quite specifically appropriate vocabulary. *Fresh in the next semester*

suggests that she's familiar with the education context as well and that's a word that's often used in that context. So I feel that this student is performing confidently given her language limitations, the technical limitations of her language, particularly the grammar. On the other hand, I suspect it would be difficult to eradicate those problems.

Appendix C

Conceptual Categories

Extracts From Verbal Reports

Set 1: Linguistic Resources: Grammar—Common Errors

Small problem with the past participle use with the passive there—"was shown" rather than *was showed*. A small error.

Again, this is not a good beginning—*they will need a monkey*—it's like Mrs. Beeton saying first "catch your ... whatever it is you need to cook." The future tense is not appropriate here. The speaker's meant to be describing an experiment that's already taken place, so it really should be in the past tense.

So she's still having trouble doing this performance in present simple to refer to a generic notion which she needs to evaluate.

So if the school cancelled—alright, we've got the conditional here and the use of the subjunctive. Good.

Again, he's using the infinitive form "be" in the conditional. Now I think he shows that he's got some idea of the conditional but does not completely understand the form of the verb "be" that needs to be used.

Even the monkey cannot be rewarded—again, the passive is there, but the actual verb choice is wrong and the sense of the sentence is wrong. It should be a conditional—"if the monkey was not rewarded" or "even if the monkey was not rewarded."

Another lack of plural noun, but she does have the use of the first conditional.

The people still start pumping the water again, "still" is wrong there, you don't need that adverb and she should've used the past tense. She doesn't seem to be able to express this in the passive. She keeps using the word "the people." I would be looking at this level for perhaps the use of the passive in this type of task.

There is some problems. So this is limitations in his command of grammar.

There's a lack of third person singular agreement—*the animal brain have.*

She is also using definite articles where they're not needed—for example *before school*, *so the students will not go to the school*—so that's a grammatical problem that she has.

Yes, here there should be a gerund, but she doesn't have gerund. She uses the infinitive

instead, saying that in order to solve the problem of land subsidence we have *just stop to pump*.

Wrong use of the modal there, although very common with Asian L1 speakers to misplace the modal—the "mights," "cans," and "mays"—both in terms of tense and meaning. "Might" here for her means "I'm not sure."

Can't form relative clauses—"the way in which students are able to choose" or something of that nature is what he wants to say.

Set 2: Linguistic Resources: Grammar—Sophistication/Range

This speaker is rather simplistic in her grammar ... She doesn't seem to have a firm grasp of simple past tense. She doesn't seem to exhibit the use of any gerunds, so her grammar is at a fairly basic level.

There's a display of variety of syntactic structure.

Again another sophisticated construction—*so just preparing them to become good consumers*—he's using again a double construction and very good use of the gerund again—*preparing them to become consumers*.

She's able to use sort of more complex construction—*is used*, for instance.

What I find amazing about this is that for something that has been prepared in such a short time, the student is actually—his performance is really presented in very complete and well structured sentences, and this is really one very good example with, you know, subordinate and main clauses, and yeah, it's perfectly structured.

Yeah, so you've got sort of quite complex sentences, which are structured and, you know, appropriately.

Again she's used the conditional so she's aware of a variety of syntactic structures that the language has.

Yeah, *academic development of any student*—further appreciation. He's able to cluster nouns together using nominalizations and prepositions to connect those in a way that a native speaker does.

Set 3: Phonology: Pronunciation—Common Problems

There are pronunciation sort of problems with "the" and "de" and things but they don't impact on understanding at all, so I think that's fine.

Propessor rather than "professor," the change of consonant there.

Again other problems with pronunciations ... is the confusion of "l" and "r" in *totally*.

She has problems with "th" at the beginning of a word, and she's pronouncing it as a "d."

Pronunciation problem—"sh" for "children"...

Pronunciation of "th" sound, so he said "mont" instead of "months."

She has severe vowel problems, pronouncing vowels, and this inaccuracy in the pronunciation of her vowels is perhaps her severest problem in terms of speech production.

And also mispronunciation—*this* or *these*—was a bit unclear because she's got a fairly staccato pronunciation, shortened vowel sounds.

Bit of interference from first language—*eet*, rather than *it want*—so a hard vowel form rather than soft.

... *nid it*—so he's given short sound where he needs to have a long sound in *need*.

Little bit of mispronunciation there—*stirred* rather than "stared"—nothing too significant.

The first vowel in "purpose"—*papurse*—and the retroflexion on the vowel in the second syllable there, make the word sound like "papers" rather than "purpose."

Again, some omission of final sound in her pronunciation, so *I thik*, rather than "I think."

That's an example which I think occurs again later on of the student having difficulty with the second and third syllable, pronouncing of the word, so for example the *ed* has not been pronounced on *adopted*.

Cutting the consonant—*groun water*—so he's cutting the end of consonants at times.

Also empathetic vowels—*studente*—extra vowels to ease consonant clusters, make comprehension more difficult.

Again, we've got L1 interference I think in the pronunciation of *use-ed*.

Appendix D

Logical Connectives

Table D1

Example D1—Task 3, Level 5

	Level 1	Level 2	Test-taker's speech	Connectives
1	Problem	Outcome	The problem is in, the problem in California is, #2 uh, the land is sinking	
2		Outcome	**and** sand is walking	additive external paratactic
3		Outcome	**and** land is move and tied together #1	additive external paratactic
4		Process	**because** of, #4 the he pum- he pump water the ground too much	consequential external hypotactic
5		Outcome	**so** the, the ground was sinking.	consequential external hypotactic
6		Process	**And** #3we pump a lot more water	additive internal paratactic
7		Outcome	**and** now is increasing so much #3	consequential internal paratactic
8		Evaluation	**and** #6 and we can uh we can stop XX we can solve the problem.	additive internal paratactic
9	Solution	Evaluation	We have to uh, we have to find another source of water #2	
10		Evaluation	**and** maybe he's can uh slow down a little bit.	consequential external paratactic

(Table continues)

Table D1 (continued)

	Level 1	Level 2	Test-taker's speech	Connectives
11	Further problem	Outcome	**Because** uh **when** this, #2 the land's sinking it can make the building #1 sinking so much too.	consequential external paratactic consequential external hypotactic
12		Process	**And** #2 **becau-** uh people #2 uh people use ah most water from a underground	additive internal paratactic unclear
13			#3 **if-** if half of- half leaking is from the ground.	

Table D2

Example D2—Task 3, Level 4

	Level 1	Level 2	Test-taker's speech	Connectives
1	Problem	process	Okay uh what happened in the San Joaquin Valley in California was that uh the valley has been overused by pumping out underground water for long long time.	
2		process	Um, mainly the water has been used uh for irrigating crops.	
3		background	Starting in uh the late eighteen- eighteen hundreds.	
4		development	**So** um, things got worse, um,	additive internal paratactic
5		development	already in the nineteen twenties the pumped out the groundwater, um, uh huge amounts of it actually	
6		outcome	**so** by nineteen sixties um the ground level had dropped about eight and a half meters which is really a lot.	additive internal paratactic
7			Um, um, **so** they,	

(Table continues)

Table D2 (continued)

	Level 1	Level 2	Test-taker's speech	Connectives
8		outcome	well it was it was a gradual drop of course, not like a you wake up and just suddenly eight and a half meters uh #2 further down	
9		comment	**but** um, #1 still it was a, a major drop	comparative external paratactic
10	solution	background	**so** uh what they thought was uh well we have to change something here	additive internal paratactic
11		process	**and** they did	additive internal paratactic
12			**and** reduced the amount of water to, that was pumped out #1 underground, um, by the early seventies I think.	additive external paratactic
13	complication	process	**But** then they had the problem that uh there was a huge drought [cough] uh I guess in the eighties uh	comparative external paratactic
14			**so** they could no longer uh use surface water	consequential external paratactic
15			**and** they had to rely again on groundwater.	consequential external paratactic
16		process	Um, they did so	
17		outcome	**and** that caused #2 again uh the ground to drop down land subsidence	additive external paratactic
18	further problem	outcome	**but** this time it was much #2 greater than the first time in comparison to the amount of time	comparative external paratactic
19		comment	**and** uh, I guess they're still trying to solve the problem by now.	

Appendix E

Descriptive Statistics for Study II—Research Question 4

Table E1

Grammatical Accuracy (1) by Proficiency Level

	Articles			Tense marking			Third-person-singular verbs		
Level	*N*	*M*	*SD*	*N*	*M*	*SD*	*N*	*M*	*SD*
1	40	69.49	25.71	30	47.89	37.43	30	69.96	34.54
2	40	73.63	22.32	36	65.19	35.06	31	64.40	41.14
3	38	70.19	17.30	40	64.12	32.28	28	70.70	36.09
4	40	75.37	17.40	39	75.20	28.31	34	78.80	32.08
5	40	84.43	17.16	40	86.58	14.87	26	91.11	19.60

Table E2

Grammatical Accuracy (2) by Proficiency Level

	Plural			Prepositions			Global accuracy		
Level	*N*	*M*	*SD*	*N*	*M*	*SD*	*N*	*M*	*SD*
1	37	66.88	33.63	39	73.28	25.49	38	16.50	24.31
2	40	58.99	34.22	39	83.07	16.58	40	21.29	27.85
3	40	73.65	27.73	40	85.26	11.73	40	20.96	19.01
4	39	81.12	16.89	40	88.16	10.26	40	30.98	22.98
5	40	94.07	9.10	40	90.71	11.99	40	50.93	21.83

Table E3

Grammatical Complexity (1)—Sentence Complexity by Proficiency Level

	T-unit complexity			Dependent-clause ratio			Verb-phrase ratio			Mean length of utterance		
Level	*N*	*M*	*SD*	*N*	*M*	*SD*	*N*	*M*	*SD*	*N*	*M*	*SD*
1	33	2.17	1.12	33	0.42	0.21	33	0.40	0.55	40	15.19	6.37
2	39	2.03	0.87	39	0.39	0.23	39	0.44	0.36	40	17.47	5.33
3	40	1.93	0.49	40	0.41	0.14	40	0.39	0.25	40	17.53	6.20
4	40	1.92	0.59	40	0.41	0.14	40	0.44	0.38	40	18.31	7.90
5	40	2.04	0.65	40	0.41	0.16	40	0.54	0.47	40	19.77	6.34

Note. T-unit complexity = the number of clauses per T-unit; Dependent-clause ratio = dependent clauses per clause; Verb-phrase ratio = verb phrases per T-unit; Mean length of utterance = number of morphemes per utterance.

Table E4

Grammatical Complexity (2)—Sophistication by Proficiency Level

		Modals		Comparatives		Passive	
Level	*N*	*M*	*SD*	*M*	*SD*	*M*	*SD*
1	40	2.39	3.07	0.38	1.33	0.34	1.12
2	40	2.90	2.59	0.20	0.55	0.76	1.06
3	40	2.40	2.21	0.16	0.47	0.68	0.88
4	40	2.56	2.40	0.21	0.38	0.88	1.17
5	40	2.63	2.23	0.19	0.49	0.72	0.84

Note. Measures are reported per 100 words.

Table E5

Use of Logical Connectives (1) by Proficiency Level

		Additive external paratactic		Comparative external paratactic		Temporal external paratactic		Consequential external paratactic		Additive internal paratactic		Comparative internal paratactic	
		And		But		Then		So		And		But	
Level	*N*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
1	16	1.33	1.73	0.39	1.56	0.001	0.001	0.35	0.95	0.39	0.90	0.17	0.68
2	16	1.48	1.24	0.38	0.99	0.001	0.001	0.56	1.07	0.32	0.68	0.04	0.17
3	14	2.22	1.84	0.45	0.59	0.29	0.52	0.57	0.76	0.39	0.65	0.31	0.47
4	12	3.24	1.55	0.62	0.85	0.17	0.30	1.24	1.24	0.40	0.63	0.27	0.35
5	12	2.48	1.46	0.27	0.44	0.23	0.29	0.82	0.67	0.49	0.68	0.23	0.37

Note. Measures are reported per 100 words.

Table E6

Use of Logical Connectives (2) by Proficiency Level

		Temporal/ external/hypotactic		Consequential/ external/hypotactic		Comparative/ internal/hypotactic	
		When		Because		Not only ... but also	
Level	*N*	*M*	*SD*	*M*	*SD*	*M*	*SD*
1	16	0.001	0.001	0.63	1.27	0.001	0.001
2	16	0.16	0.48	1.26	1.24	0.001	0.001
3	14	0.38	1.06	1.54	1.73	0.001	0.001
4	12	0.14	0.50	0.79	0.96	0.17	0.60
5	12	0.19	0.38	1.29	1.39	0.05	0.16

Note. Measures are reported per 100 words.

Table E7

Vocabulary Use (1) by Proficiency Level

		Word-Token		Word-Type		Type-token	
Level	*N*	*M*	*SD*	*M*	*SD*	*M*	*SD*
1	40	55.68	18.86	38.02	12.94	0.69	0.12
2	40	69.92	18.76	42.73	9.78	0.63	0.11
3	40	86.87	20.08	49.05	12.97	0.57	0.08
4	40	100.08	22.62	56.39	14.04	0.56	0.08
5	40	118.09	22.64	66.04	14.55	0.56	0.07

Table E8

Vocabulary Use (2) by Proficiency Level

		K1 (%)		K2 (%)		AWL (%)		Off-list (%)	
Level	*N*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
1	40	87.18	6.54	5.02	4.43	3.54	3.38	4.26	5.31
2	40	88.42	6.17	4.80	3.06	2.44	1.94	4.66	4.50
3	40	88.89	4.46	4.21	2.83	2.80	2.43	4.12	3.10
4	40	88.30	3.76	4.05	2.45	3.56	1.94	3.83	2.35
5	40	86.78	3.87	4.07	2.03	4.27	2.26	4.87	3.44

Note. K1 = the most frequent 1,000 words of English; K2 = the second most frequent 1,000 words of English; AWL = words from the Academic Word List; Off-list = all remaining words.

Table E9

Pronunciation by Proficiency Level

	Level 1		Level 2		Level 3		Level 4		Level 5	
	N = 14		*N* = 16		*N* = 16		*N* = 17		*N* = 16	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Word-level analysis										
Meaningful words on target	8.68	1.05	8.44	1.01	8.56	0.69	8.98	0.58	9.06	0.58
Meaningful words, but marginally nontarget	0.13	0.21	0.14	0.14	0.16	0.20	0.14	0.18	0.09	0.11
Marginally nontarget morpho. ending	0.11	0.16	0.08	0.14	0.10	0.18	0.12	0.18	0.02	0.05
Marginally nontarget stress, V-fullness/ reduction	0.02	0.09	0.06	0.10	0.06	0.09	0.02	0.06	0.07	0.10
Meaningful words, but clearly nontarget	0.12	0.23	0.23	0.38	0.13	0.16	0.16	0.19	0.03	0.07
Clearly nontarget morpho. ending	0.001	0.001	0.05	0.11	0.01	0.06	0.01	0.04	0.02	0.05
Clearly nontarget stress, V-fullness/ reduction	0.12	0.23	0.18	0.33	0.12	0.16	0.15	0.20	0.01	0.05
Nonmeaningful words	1.06	0.92	1.19	0.92	1.16	0.66	0.72	0.48	0.82	0.61
*Um*s, *er*s, etc.	0.88	0.73	0.74	0.62	0.86	0.60	0.59	0.45	0.69	0.52
Incomprehensible words	0.05	0.13	0.09	0.19	0.05	0.09	0.04	0.09	0.02	0.08
Words not completed	0.14	0.31	0.35	0.46	0.25	0.31	0.09	0.13	0.11	0.15
Subword-level analysis										
On target syllables	8.25	0.93	8.03	0.85	8.64	0.68	9.08	0.53	9.46	0.30
Marginally nontarget syllables	0.31	0.29	0.35	0.24	0.29	0.29	0.22	0.21	0.12	0.10
Marginally whispered/ swallowed etc.	0.11	0.15	0.10	0.14	0.06	0.10	0.10	0.16	0.06	0.07

(Table continues)

Table E9 continued

	Level 1		Level 2		Level 3		Level 4		Level 5	
	N = 14		*N* = 16		*N* = 16		*N* = 17		*N* = 16	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Marginally nontarget <th>	0.06	0.11	0.04	0.09	0.04	0.08	0.04	0.06	0.03	0.05
Full consonant, but marginally nontarget	0.07	0.14	0.14	0.18	0.11	0.18	0.02	0.05	0.03	0.06
Marginally nontarget full vowel	0.07	0.16	0.08	0.09	0.08	0.12	0.06	0.11	0.01	0.03
Clearly nontarget-like syllables	1.27	0.95	1.39	0.73	1.01	0.58	0.65	0.54	0.42	0.30
Clearly whispered/ swallowed, etc.	0.47	0.47	0.53	0.39	0.27	0.23	0.19	0.19	0.21	0.22
Full consonant, but clearly nontarget	0.29	0.30	0.22	0.21	0.14	0.13	0.06	0.07	0.04	0.06
Clearly nontarget <th>	0.41	0.46	0.54	0.49	0.53	0.38	0.32	0.36	0.13	0.15
Clearly nontarget full vowel	0.10	0.18	0.11	0.18	0.08	0.14	0.08	0.14	0.03	0.06
Epenthetic syllable, clear	0.16	0.10	0.18	0.11	0.18	0.08	0.14	0.08	0.14	0.03
Epenthetic syllable, marginal	0.02	0.06	0.08	0.11	0.05	0.09	0.02	0.04	0.01	0.02

Note. All data reported in this table are frequency data; word level = per 10 words; subword level = per 10 syllables in meaningful words.

Table E10

Intonation by Proficiency Level

	Native like ←			→ Nonnative like	
	Many English-like	Many near English-like	Many non-English-like	Relatively few	Total
Level 1	1	0	7	6	14
Level 2	0	0	10	6	16
Level 3	2	2	5	7	16
Level 4	3	4	9	0	16
Level 5	7	5	4	0	16
Total	13	11	35	19	78

Table E11

Rhythm by Proficiency Level

	Native-like ←			→ Nonnative-like	
	Stress-timed	Variable	Unclear	Syllable-timed	Total
Level 1	3	0	6	5	14
Level 2	3	2	4	7	16
Level 3	6	1	4	5	16
Level 4	6	4	4	2	16
Level 5	10	4	2	0	16
Total	28	11	20	19	78

Table E12

Fluency by Proficiency Level

		Filled pauses		Unfilled pauses		Total pause time		Repairs		Speech rate		Mean length of run	
Level	*N*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
1	39	4.62	4.88	5.98	5.44	24.78	19.71	4.35	2.83	1.32	0.44	26.81	19.32
2	40	6.07	4.63	5.13	2.83	18.68	11.70	4.45	3.07	1.66	0.44	22.85	12.25
3	40	6.06	4.58	3.93	2.43	11.22	7.79	4.39	2.63	2.02	0.45	20.84	11.22
4	39	6.61	4.91	3.04	2.84	7.79	8.05	4.49	2.10	2.36	0.46	21.30	11.85
5	40	6.64	5.64	1.49	1.83	3.81	4.68	3.46	2.44	2.83	0.50	22.80	10.16

Note. The numbers of filled pauses, unfilled pauses, and repairs reported in the table are frequency data (per 60 s).

Table E13

Quantity of Discourse by Proficiency Level

		T-unit		Clause	
Level	*N*	*M*	*SD*	*M*	*SD*
1	40	8.26	3.06	13.39	7.39
2	40	8.35	1.97	15.68	6.92
3	40	8.81	1.92	17.88	4.73
4	40	8.80	2.10	18.99	6.77
5	40	8.74	1.98	20.07	6.90

Appendix F

Descriptive Statistics for Study II—Research Question 5

Table F1

Grammatical Accuracy (1) by Task and Task Type

	Articles			Tense marking			Third-person-singular verbs		
Task	*N*	*M*	*SD*	*N*	*M*	*SD*	*N*	*M*	*SD*
1	40	76.78	21.27	36	80.37	32.60	31	81.4	28.36
2	39	64.30	25.18	38	63.20	35.66	30	65.00	42.29
3	39	80.36	12.38	37	56.27	31.08	18	76.64	34.24
4	40	73.37	22.02	38	76.28	24.95	35	74.86	35.50
5	40	77.92	17.86	36	68.57	32.11	35	75.71	31.14
Independent	79	70.62	23.97	74	71.55	35.05	61	73.33	36.53
Integrated	119	77.35	18.02	111	67.11	30.38	88	75.57	33.18

Table F2

Grammatical Accuracy (2) by Task and Task Type

	Plural			Preposition			Global accuracy		
Task	*N*	*M*	*SD*	*N*	*M*	*SD*	*N*	*M*	*SD*
1	38	75.7	28.58	38	84.61	16.77	39	28.22	31.01
2	40	76.6	23.74	40	81.61	18.54	40	28.33	28.99
3	38	71.6	29.06	40	82.87	20.63	40	32.24	28.74
4	40	74.36	31.85	40	83.87	13.91	40	27.87	20.96
5	40	76.75	30.17	40	87.84	14.72	39	24.51	20.35
Independent	78	76.16	26.04	78	83.07	17.65	79	28.28	29.81
Integrated	118	74.28	30.22	120	84.86	16.69	119	28.23	23.7

Table F3

Grammatical Complexity (1)—Sentence Complexity by Task and Task Type

	T-unit complexity			Dependent-clause ratio			Verb-phrase ratio			Mean length of utterance		
Task	*N*	*M*	*SD*	*N*	*M*	*SD*	*N*	*M*	*SD*	*N*	*M*	*SD*
1	35	2.27	0.87	35	0.46	0.19	35	0.42	0.35	40	20.30	7.21
2	39	2.18	0.70	39	0.44	0.15	39	0.44	0.52	40	18.54	8.83
3	39	1.51	0.37	39	0.28	0.13	39	0.43	0.24	40	14.18	3.96
4	40	2.02	0.88	40	0.42	0.18	40	0.40	0.38	40	17.29	5.92
5	39	2.10	0.65	39	0.45	0.18	39	0.52	0.50	40	17.96	4.48
Independent	74	2.22	0.78	75	0.45	0.17	77	0.44	0.48	80	19.42	8.06
Integrated	118	1.88	0.71	119	0.39	0.19	118	0.45	0.39	120	16.47	5.09

Note. T-unit complexity = the number of clauses per T-unit; Dependent-clause ratio = dependent clauses per clause; Verb-phrase ratio = verb phrases per T-unit; Mean length of utterance = number of morphemes per utterance.

Table F4

Adjusted Means for Sentence Complexity—T-Unit Complexity by Task Type

			95% confidence interval	
Task	Mean	Standard error	Lower bound	Upper bound
Independent	2.11	0.09	1.93	2.29
Integrated	1.96	0.07	1.83	2.10

Table F5

Adjusted Means for Sentence Complexity—Dependent-Clause Ratio by Task Type

			95% confidence interval	
Task	Mean	Standard error	Lower bound	Upper bound
Independent	0.41	0.02	0.37	0.45
Integrated	0.41	0.02	0.38	0.44

Table F6

Adjusted Means for Sentence Complexity—Verb-Phrase Ratio by Task Type

			95% confidence interval	
Task	Mean	Standard error	Lower bound	Upper bound
Independent	0.36	0.05	0.26	0.46
Integrated	0.50	0.04	0.42	0.58

Table F7

Adjusted Means for Sentence Complexity—Mean Length of Utterance by Task Type

			95% confidence interval	
Task	Mean	Standard error	Lower bound	Upper bound
Independent	17.83	0.67	16.52	19.15
Integrated	17.65	0.54	16.59	18.72

Table F8

Grammatical Complexity (2)—Sophistication by Task and Task Type

		Modal		Comparative		Passive	
Task	*N*	*M*	*SD*	*M*	*SD*	*M*	*SD*
1	40	4.48	2.51	0.17	0.52	0.78	1.32
2	40	3.96	2.19	0.08	0.35	0.77	1.21
3	40	0.88	1.29	0.20	0.76	0.75	0.89
4	40	2.35	1.76	0.21	0.48	0.59	0.84
5	40	1.21	2.34	0.48	1.19	0.49	0.79
Independent	80	4.22	2.35	0.13	0.44	0.77	1.26
Integrated	120	1.48	1.94	0.30	0.86	0.61	0.84

Note. Descriptive statistics are reported from frequency data (per 100 words).

Table F9

Use of Logical Connectives (1) by Task

		Additive external paratactic		Comparative external paratactic		Temporal external paratactic		Consequential external paratactic		Additive internal paratactic		Comparative internal paratactic	
		And		But		Then		So		And		But	
Task	*N*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
2	30	1.44	1.57	0.46	1.34	0.00	0.00	0.36	0.72	0.27	0.60	0.04	0.24
3	40	2.53	1.63	0.38	0.61	0.22	0.38	0.91	1.09	0.49	0.77	0.31	0.53

Note. Descriptive statistics are reported from frequency data (per 100 words).

Table F10

Use of Logical Connectives (2) by Task

		Temporal external hypotactic		Consequential external hypotactic		Comparative internal hypotactic	
		When		Because		Not only ... but also	
Task	*N*	*M*	*SD*	*M*	*SD*	*M*	*SD*
2	30	0.19	0.79	1.50	1.45	0.07	0.38
3	40	0.15	0.38	0.79	1.20	0.01	0.09

Note. Descriptive statistics are reported from frequency data (per 100 words).

Table F11

Vocabulary Use (1) by Task and Task Type

		Word-token		Word-type		Type-token ratio	
Task	*N*	*M*	*SD*	*M*	*SD*	*M*	*SD*
1	40	96.54	35.41	60.40	18.01	0.65	0.11
2	40	92.70	29.05	59.81	14.48	0.67	0.10
3	40	77.02	26.11	41.15	12.42	0.55	0.09
4	40	87.98	27.50	46.30	12.49	0.53	0.07
5	40	76.40	26.93	44.58	13.01	0.60	0.09
Independent	80	94.62	32.24	60.11	16.24	0.66	0.10
Integrated	120	80.47	27.16	44.01	12.72	0.56	0.09

Note. Descriptive statistics for word-token and word-type data are reported from frequency data (per 60 s).

Table F12

Adjusted Means for Type-Token Ratio by Task and Task Type

			95% confidence interval	
Task	Mean	Standard error	Lower bound	Upper bound
Independent	0.62	0.01	0.60	0.64
Integrated	0.59	0.01	0.57	0.60

Table F13

Vocabulary Use (2) by Task and Task Type

		K1 (%)		K2 (%)		AWL (%)		Off-list (%)	
Task	*N*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
1	40	90.30	4.44	3.08	2.21	3.72	2.49	2.66	3.00
2	40	90.96	4.16	2.70	2.20	2.88	2.74	3.45	2.48
3	40	84.66	4.20	6.09	2.25	3.32	1.86	6.16	3.64
4	40	87.23	3.97	7.49	2.92	3.33	2.67	1.97	1.71
5	40	86.43	5.81	2.79	2.12	3.36	2.71	7.51	4.65
Independent	80	90.63	4.29	2.89	2.20	3.30	2.63	3.06	2.77
Integrated	120	86.10	4.81	5.46	3.14	3.34	2.43	5.21	4.24

Note. K1 = the most frequent 1,000 words of English; K2 = the second most frequent 1,000 words of English; AWL = words from the Academic Word List; Off-list = all remaining words.

Table F14

Pronunciation by Task and Task Type

	Task 2		Task 3	
	N = 38		*N* = 41	
	M	*SD*	*M*	*SD*
Word level				
Meaningful words on target	8.82	0.78	8.68	0.85
Meaningful words, but marginally nontarget	0.16	0.18	0.11	0.16
Marginally nontarget morpho. ending	0.10	0.17	0.07	0.13
Marginally nontarget stress, V-fullness/reduction	0.05	0.09	0.04	0.09
Meaningful words, but clearly nontarget	0.16	0.28	0.11	0.17
Clearly nontarget morpho. ending	0.02	0.06	0.01	0.06
Clearly nontarget stress, V-fullness/reduction	0.14	0.25	0.09	0.17

(Table continues)

Table F14 (continued)

	Task 2		Task 3	
	N = 38		*N* = 41	
	M	*SD*	*M*	*SD*
Nonmeaningful words	0.86	0.58	1.10	0.84
Ums, ers, etc	0.68	0.55	0.81	0.61
Incomprehensible words	0.08	0.15	0.02	0.07
Words not completed	0.10	0.16	0.27	0.38
Subword level				
On target syllables	8.89	0.88	8.54	0.81
Marginally nontarget syllables	0.24	0.27	0.27	0.22
Marginally whispered/swallowed etc.	0.09	0.15	0.08	0.10
Marginally nontarget full realization of <th>	0.07	0.13	0.08	0.15
Full consonant, but marginally nontarget	0.02	0.05	0.02	0.05
Marginally nontarget quality of a full vowel	0.06	0.12	0.05	0.10
Clearly nontarget syllables	0.77	0.69	1.08	0.74
Clearly whispered/swallowed etc.	0.24	0.32	0.41	0.34
Full consonant, but clearly nontarget	0.11	0.15	0.17	0.23
Clearly nontarget full realization of <th>	0.34	0.41	0.43	0.40
Clearly nontarget quality of a full vowel	0.09	0.15	0.07	0.14
Epenthetic syllable, clear	0.07	0.17	0.06	0.21
Epenthetic syllable, marginal	0.03	0.08	0.03	0.07

Note. Word level = per 10 words; Subword level = per 10 syllables in meaningful words.

Table F15

Intonation by Task

	Native like ←			→ Nonnative like	
Task	Many English-like	Many near English-like	Many non-English-like	Relatively few	Total
2	9	4	18	7	38
3	4	7	17	12	40

Table F16

Rhythm by Task

Task	Native like ← Stress-timed	Variable	Unclear	→ Nonnative like Syllable-timed	Total
2	24	4	3	7	38
3	4	7	17	14	40

Table F17

Fluency by Task

		Filled pauses		Unfilled pauses		Total pause time		Repairs		Speech rate		Mean length of run	
Task	*N*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
1	40	5.32	5.37	3.29	3.46	10.00	11.13	4.33	3.33	2.20	0.81	32.07	12.88
2	40	7.03	5.33	4.88	5.57	15.46	17.81	4.08	2.48	2.20	0.69	28.82	13.77
3	38	6.56	5.12	4.66	2.97	14.59	12.86	4.49	2.72	1.83	0.64	12.54	4.10
4	40	5.49	4.12	3.80	2.72	10.32	10.88	4.32	2.35	2.13	0.69	20.25	14.77
5	40	5.62	4.71	2.94	2.39	15.92	14.39	3.94	2.27	1.82	0.56	20.35	8.47
Indepen-dent	80	6.18	5.38	4.08	4.67	12.73	15.01	4.20	2.92	2.20	0.75	30.44	13.35
Inte-grated	120	5.87	4.61	3.80	2.77	13.61	12.91	4.20	2.45	1.93	0.64	17.80	10.73

Note. The numbers of filled pauses, unfilled pauses, and repairs reported in the table represent frequency data (per 60 s).

Table F18

Quantity of Discourse (by Task)

		T-units		Clauses	
Task	*N*	*M*	*SD*	*M*	*SD*
1	40	6.59	3.50	18.41	8.78
2	40	7.88	2.10	17.97	7.18
3	40	9.17	0.93	13.51	2.70
4	40	9.66	0.57	18.18	6.66
5	40	9.66	0.66	17.85	7.04
Independent	80	8.30	2.56	18.19	7.97
Integrated	120	8.78	1.98	16.49	6.13

Note. Numbers of T-units and clauses represent frequency data (per 10 utterances).

Appendix G

ANOVA/ANCOVA Tables for Study II

Table G1

Comparison of Grammatical Accuracy

Source	Sum of squares	*Df*	Mean square	*F*	*p*	Effect size
Articles						
Level	5,599.24	4	1,399.81	3.41	0.001	0.07
Task type	2,177.33	1	2,177.33	5.30	0.02	0.03
Level by task type	179.73	4	44.93	0.11	0.98	0.00
Error	77,215.43	188	410.72			
Tense-marking						
Level	27,531.56	4	6,882.89	7.45	0.001	0.15
Task type	1,053.46	1	1,053.46	1.14	0.29	0.01
Level by task type	544.15	4	136.04	0.15	0.96	0.00
Error	161,737.22	175	924.21			
Third-person-singular verbs						
Level	13,546.61	4	3,386.65	3.01	0.02	0.08
Task type	192.59	1	192.59	0.17	0.68	0.00
Level by task type	7,425.31	4	1,856.33	1.65	0.17	0.05
Error	156,266.96	139	1,124.22			
Plural nouns						
Level	26,801.22	4	6,700.30	9.58	0.001	0.17
Task type	143.27	1	143.27	0.21	0.65	0.00
Level by task type	281.10	4	70.28	0.10	0.98	0.00
Error	114,373.38	173	661.12			
Prepositions						
Level	7,694.58	4	1,923.65	7.42	0.001	0.14
Task type	195.31	1	195.31	0.75	0.39	0.00
Level by task type	1,313.34	4	328.34	1.27	0.29	0.03
Error	48,734.31	188	259.23			
Global errors						
Level	30,082.12	4	7,520.53	13.51	0.001	0.22
Task type	0.00	1	0.001	0.000	1.00	0.00
Level by task type	708.29	4	177.07	0.32	0.87	0.01
Error	104,679.06	188	556.80			

Table G2

Comparison of Grammatical Complexity (1)—Sentence Complexity

Source	Sum of squares	*Df*	Mean square	*F*	*p*	Effect size
T-unit complexity						
Utterance	8.65	1	8.65	17.00	0.001	0.09
Level	1.93	4	0.48	0.95	0.44	0.02
Task type	0.79	1	0.79	1.54	0.22	0.01
Level by task type	2.39	4	0.60	1.17	0.32	0.03
Error	91.10	179	0.51			
Dependent clause ratio						
Utterance	0.66	1	0.66	21.58	0.001	0.11
Level	0.17	4	0.04	1.40	0.24	0.03
Task type	0.00	1	0.00	0.00	0.97	0.00
Level by task type	0.05	4	0.01	0.37	0.83	0.01
Error	5.57	181	0.03			
Verb phrase complexity						
Utterance	2.66	1	2.66	15.87	0.001	0.08
Level	2.35	4	0.59	3.50	0.01	0.07
Task type	0.69	1	0.69	4.11	0.04	0.02
Level by task type	0.39	4	0.10	0.58	0.68	0.01
Error	30.49	182	0.17			
Mean length of utterances (MLU)						
Utterance	1,677.85	1	1,677.85	53.09	0.001	0.22
Level	1,362.07	4	340.52	10.77	0.001	0.19
Task type	1.29	1	1.29	0.04	0.84	0.00
Level by task type	116.93	4	29.23	0.93	0.45	0.02
Error	5,910.02	187	31.60			

Table G3

Comparison of Grammatical Complexity (2)—Sophistication

Source	Sum of squares	*Df*	Mean square	*F*	*p*	Effect size
Modal						
Level	9.10	4	2.28	0.50	0.73	0.01
Task type	360.38	1	360.38	79.79	0.001	0.30

(Table continues)

Table G3 (continued)

Source	Sum of squares	*Df*	Mean square	*F*	*p*	Effect size
Level by task type	19.33	4	4.83	1.07	0.37	0.02
Error	858.19	190	4.52			
Comparative						
Level	0.82	4	0.21	0.38	0.82	0.01
Task type	1.37	1	1.37	2.54	0.11	0.01
Level by task type	0.51	4	0.13	0.24	0.92	0.01
Error	102.70	190	0.54			
Passive						
Level	6.50	4	1.63	1.56	0.19	0.03
Task type	1.29	1	1.29	1.23	0.27	0.01
Level by task type	3.57	4	0.89	0.85	0.49	0.02
Error	198.56	190	1.05			

Table G4

Comparison of Vocabulary Use (1)

Source	Sum of squares	*Df*	Mean square	*F*	*p*	Effect size
Word-token						
Level	95,006.35	4	23,751.59	62.32	0.001	0.57
Task type	9,612.42	1	9,612.42	25.22	0.001	0.12
Level by task type	1,237.19	4	309.30	0.81	0.52	0.02
Error	72,408.65	190	381.10			
Word-type						
Level	20,008.33	4	5,002.08	47.88	0.001	0.50
Task type	12,440.28	1	12,440.28	119.08	0.001	0.39
Level by task type	476.85	4	119.21	1.14	0.34	0.02
Error	19,849.94	190	104.47			
Type-token ratio						
Length of speech	0.20	1	0.20	35.53	0.001	0.16
Level	0.30	4	0.07	13.23	0.001	0.22
Task type	0.02	1	0.02	4.42	0.04	0.02
Level by task type	0.01	4	0.00	0.32	0.87	0.01
Error	1.07	189	0.01			

Table G5

Comparison of Vocabulary Use (2)

Source	Sum of squares	*Df*	Mean square	*F*	*p*	Effect size
K1 (%)						
Level	151.06	4	37.77	1.90	0.11	0.04
Task type	983.39	1	983.39	49.36	0.001	0.21
Level by task type	292.30	4	73.07	3.67	0.01	0.07
Error	3,785.64	190	19.92			
K2 (%)						
Level	20.61	4	5.15	0.68	0.60	0.01
Task type	315.67	1	315.67	41.91	0.001	0.18
Level by task type	89.06	4	22.27	2.96	0.02	0.06
Error	1,431.21	190	7.53			
AWL (%)						
Level	96.03	4	24.01	3.99	0.001	0.08
Task type	0.06	1	0.06	0.01	0.92	0.00
Level by task type	24.31	4	6.08	1.01	0.40	0.02
Error	1,143.08	190	6.02			
Off-list (%)						
Level	21.39	4	5.35	0.39	0.82	0.01
Task type	223.60	1	223.60	16.09	0.001	0.08
Level by task type	78.53	4	19.63	1.41	0.23	0.03
Error	2,640.01	190	13.90			

Note. K1 = the most frequent 1,000 words of English; K2 = the second most frequent 1,000 words of English; AWL = words from the Academic Word List; Off-list = all remaining words.

Table G6

Comparison of Phonological Features

Source	Sum of squares	*Df*	Mean square	*F*	*p*	Effect size
Meaningful words						
Level	4.60	4	1.15	1.75	0.15	0.09
Task type	0.41	1	0.41	0.63	0.43	0.01
Level by task type	1.49	4	0.37	0.57	0.69	0.03
Error	45.37	69	0.66			

(Table continues)

Table G6 (continued)

Source	Sum of squares	*Df*	Mean square	*F*	*p*	Effect size
Meaningful words						
Level	21.29	4	5.32	11.49	0.001	0.40
Task type	2.19	1	2.19	4.72	0.03	0.06
Level by task type	0.96	4	0.24	0.52	0.72	0.03
Error	31.97	69	0.46			

Table G7

Comparison of Fluency Measures

Source	Sum of squares	*Df*	Mean square	*F*	*p*	Effect size
Number of filled pauses						
Level	72.01	4	18.00	0.75	0.56	0.02
Task type	4.59	1	4.59	0.19	0.66	0.00
Level by task type	147.48	4	36.87	1.53	0.19	0.03
Error	4,572.31	190	24.07			
Number of unfilled pauses						
Level	537.56	4	134.39	12.19	0.001	0.20
Task type	3.87	1	3.87	0.35	0.55	0.00
Level by task type	46.40	4	11.60	1.05	0.38	0.02
Error	2,094.93	190	11.03			
Total pause time						
Level	11,339.45	4	2,834.86	20.62	0.001	0.30
Task type	37.07	1	37.07	0.27	0.60	0.00
Level by task type	83.11	4	20.78	0.15	0.96	0.00
Error	26,123.70	190	137.49			
Repair						
Level	28.35	4	7.09	0.99	0.41	0.02
Task type	0.00	1	0.00	0.00	0.99	0.00
Level by task type	1.01	4	0.25	0.04	1.00	0.00
Error	1,359.72	190	7.16			
Speech rate						
Level	55.07	4	13.77	71.32	0.001	0.60
Task type	3.33	1	3.33	17.27	0.001	0.08
Level by task type	1.20	4	0.30	1.56	0.19	0.03
Error	36.49	189	0.19			

(Table continues)

Table G7 (continued)

Source	Sum of squares	*Df*	Mean square	*F*	*p*	Effect size
Mean length of run						
Level	893.78	4	223.45	1.60	0.18	0.03
Task type	7,612.60	1	7,612.60	54.54	0.001	0.23
Level by task type	469.82	4	117.46	0.84	0.50	0.02
Error	26240.43	188	139.58			

Table G8

Comparison of Quantity of Discourse Measures

Source	Sum of squares	*Df*	Mean square	*F*	*p*	Effect size
T-unit						
Utterance	656.11	1	656.11	237.77	0.001	0.56
Level	192.26	4	48.06	17.42	0.001	0.27
Task type	57.08	1	57.08	20.68	0.001	0.10
Level by task type	12.74	4	3.19	1.15	0.33	0.02
Error	516.02	187	2.76			
Clauses						
Utterance	711.59	1	711.59	70.99	0.001	0.28
Level	1,153.23	4	288.31	28.76	0.001	0.38
Task type	133.73	1	133.73	13.34	0.001	0.07
Level by task type	56.36	4	14.09	1.41	0.23	0.03
Error	1,874.42	187	10.02			

Appendix H

Examples of Sentence Complexity at Different Proficiency Levels

Example H1—Level 1 (Task 4)

(1) Human brain and monkey's brain have some similarity #2 **(1 T-unit, 1 clause)**

(2) and uh there are some theory that #2 uh um, #2 this similarity make #2 er, relate to the language #2 ss-, #11 **(1 T-unit, 2 clauses)**

(2 T-units, 3 clauses) (Rhesus_1-130037)

Example H2—Level 2 (Task 4)

(1) For this experiment in the class of this professor they will need the monkey that makes eh different things, like different set [?] the certain colors or shapes and #2 eh #2 maybe with- with objects **(1 T-unit, 2 clauses)**

(2) the the monkey will #2 will #2 will differentiate eh certain shapes and the numbers eh, #1 he would um, #2 **(1 T-unit, 1 clause)**

(3) the monkey #7 the monkey should #6 the monkey should uh, #2 should have the ability of differentiate eh shapes and objects **(1 T-unit, 1 clause)**

(4) and they- it will work with numbers too. **(1 T-unit, 1 clause)**

(5) To, in order to see if the monkey, if the monkey can eh, can order the, the numbers in uh, #3 in uh good way **(Fragment, 1 clause)**

(4 T-units, 6 clauses) (Rhesus_2-320028)

Example H3—Level 3 (Task 4)

(1) The experiment is about uh uh how to teach a monkey to count numbers. (1 T-unit, 1 clause)

(2) They put the object different ab- the num- different number of objects and then ask the monkey to touch the object. **(1 T-unit, 2 clauses)**

(3) First they touch the #2 one object. **(1 T-unit, 1 clause)**

and sec- and then they touch the twos objects, then three objects and go and so on. **(1 T-unit, 2 clauses)**

(4) And uh if they touch the right thing they will be get reward #1 and um. **(1 T-unit, 2 clauses)**

(5) #2 Uh, I me- if they touch the, touch the object in the right order they will get reward and um. **(1 T-unit, 2 clauses)**

(6) Just want to see #1 wha- uh whats the different between human brain and uh the monkey's brain. **(1 T-unit, 2 clauses)**

(7) and um #2 the experiment shows the numeracy for human the mean brain for human and uh monkey are different. (1 T-unit, 2 clauses)

(8 T-units, 15 clauses) (Rhesus_3-130067)

Example H4—Level 4 (Task 4)

(1) The experiment about which the two persons talk about was applied to monkeys. **(1 T-unit, 2 clauses)**

(2) The- they were shown a group of images like for example two circles, three flowers, four squares**. (1 T-unit, 1 clause)**.

(3) and the- they were meant to point to these figures in order eh. **(1 T-unit, 1 clause)**

(4) This was in order to prove the monkeys ability to discern about numbers. **(1 T-unit, 1 clause.**

(5) If they were- if they pointed in the correct order that's one, two, three, four, #1 they were given a reward. **(1 T-unit, 3 clauses)**.

(6) Um, if they did it in the right order eh the result was that they had the concept of numbers. **(1 T-unit, 3 clauses)**.

(7) This is called numeracy. **(1 T-unit, 1 clause)**.

(8) They were also able to recognize larger and and in different order like for example, first the seven then the ten, before goes the three, **(1 T-unit, 2 clause)**.

(9) and this kind of stuff that led a- led the investigators towards believe that the monkeys have an ability, have a language, uh because in order to have the numeracy they need to have a language. **(1 T-unit, 4 clauses)**.

(9 T-units, 17 clauses). (Rhesus_4-320003)

Example H5—Level 5 (Task 4)

(1) Well, the experiment uh consisted in that monkey- they would put four different images to monkeys right. **(1 T-unit, 2 clauses)**.

(2) One would be like one, **(1 T-unit, 1 clause)**.

(3) the second would be like describe two objects, the third three objects and the fourth four objects. **(1 T-unit, 1 clause).**

(4) So the monkeys would usually go- the seventy-five- seventy-five percent of the monkeys wan- first with one object, then with two objects, three objects and four objects. **(1 T-unit, 1 clause)**.

(5) So what they concluded about this is that there i- there really is some concept of numeracy in in monkeys right. **(1 T-unit, 3 clauses)**.

(6) Then they did this experiment with uh numbers five through nine. **(1 T-unit, 1 clause)**.

(7) So again seventy-five percent of the monkeys got it okay **(1 T-unit, 1 clause)**.

(8) and they didn't get rewarded, like they did the first time. **(1 T-unit, 2 clauses)**.

(9) So as I mentioned they concluded that there is some concept in numeracy in monkeys **(1 T-unit, 3 clauses)**.

(10) and it's quite impressive because uh not even babies, one year old, have this concept

(11) but these monkeys do **(1 T-unit, 3 clauses)**

(12) so #1 that's it. **(1 T-unit, 1 clause)**.

(12 T-units, 19 clauses). (Rhesus_5-320071)

Appendix I

Comparative Discourse Quality

Performances of Two Test-Takers on Tasks 2 and 3

Example I1—Task 2, Level 4 (Test-Taker A)

Structure	Test-taker's speech
Introduction	A complete and thorough #1 overview #2 uh #5 uh #2 of uh students education really important
	and one of those is uh music and art courses,
Opinion	and they should never be cancelled.
Reason for opinion	Personally I'm in the music #1 in some music #1 courses and the choir
	and I've loved it there
	and I think people should need not only to develop their their #2 mathematical ability but also their more #2 creative, and kind of sensitive sides.

Example I2—Task 3, Level 4 (Test-Taker A)

Level 1	Level 2	Test-taker's speech
Problem	Process	Well XX the problem that occurred in the valley in California #1 is that they started pumping out water eh at a major rate
	Outcome	so the valley started to, #2 to go down, to to sink, the whole valley at a really #2 fast rate.
	Outcome	And also the the overuse of th- pumping out water caused land subsidence.
Solution	Process	Um, #2 so what they thought about was to to get water from the, from a surface and pump it in so, to stop this, #2 the sinking.
Complication		And, #3 and the thing was that the- the- there was a drought
	Outcome	and they couldn't get any water, the water th- th- that they were thinking they could get to pump in,
		so, just continue to, #3 well, they couldn't #1 pump in #2 water
	Outcome	and the valley continued to sink.

Example I3—Task 2, Level 5 (Test-Taker B)

Structure	Test-taker's speech
Reason 1 for opinion 1	Training in art music is indispensable to the overall growth of a student.
Opinion 1	Therefore I think #1 it is very unfortunate that music and art courses might be cancelled because of budgetry constraints.
Reason 2 for opinion1	The appreciation of music is something that one must learn as a child and leads to a whole lot of benefits over one's lifetime.
Reason 3 for opinion 1	Similarly the appreciation of art is something which increases your perception
	and uh #2 may be of great importance to you if you become either an engineer or an architect to have greater perception and visualisation skills.
Reason 4 for opinion 1	Similarly, music is what in some sense differentiates man from beast.
Conclusion	It cannot be removed from one's learning.
Reason 1 for opinion1	Training in art music is indispensable to the overall growth of a student.

Example I4—Task 3, Level 5 (Test-Taker B)

Level 1	Level 2	Test-taker's speech
Introduction	Process	The San Juaquin Valley, #1 prz- presented as a place where land subsidence occurred.
Problem	Process	The San Juaquin Valley, located in California, was using #1 groundwater from the late eighteen-eighties.
	Process	Now, there was heavy pumping of water for both irrigation and other purposes #1 in this valley.
	Process	but it occurred over a long period of time.
Solution	Process	So in order to mitigate this problem in the nineteen-seventies, San Juaquin Valley reduced pumping of water and increased the use of surface water;.
	Process	Uh what they did uh they just decided to import the water from other area, surface water.
Complication	Process	however, the problem of land subsidence reappeared in the nineteen-nineties because of the drought in California.
	Process	And this made #1 people start using groundwater again.
	Outcome	And it was even a huger problem now because land levels-groundwater levels ah sunk by much greater than the seventies
		and the land level sunk greatly too.

Appendix J

Examples of Modal Usage in Specific Task Performances

Example J1 —Task 1 (School, Independent), Level 1

More recently there have been proposals in some cities that high school students **should** attend class 12 months a years. Ah if the plans was adapt there it **will** make the #4 students #6 they **can** make the students #6 make the stu- they know more more about the economy and then more familiarize with the city they #4 On the other hand,

Example J2—Task 1 (School, Independent), Level 4

If they adopt such a policy then there **will** be no blocks of uh time that people can or the students **can** have holidays and X indeed whole community because their parents are also most of the time mm connected to their children. So tourism er **will** slow down in these kind of places and eventually children **will** be bored because of the you know continuous school time. Also parents **will** be bored because they **will** not **be able to** leave the city for uh more than couple of days which are weekends.

Example J3 —Task 2 (Music, Independent), Level 2

I think musics and arts course is very important for the school. If school budgets be cut it #1 **should** not be improper for the student to uh have a chance to study in music and arts. Um, the music and arts course #2 is a basic for some #3 uh some department for the student who want to further study in this field. #6 I think the school **should** be to give a more chance for the student #2 to study in many fields in, in their paths that they **would** like to study.

Example J4 —Task 2 (Music, Independent), Level 5

I think uh art and music are very important because uh they make- you are **supposed to** have a more rounded uh education and basically know about what's going on around you and music and art are a huge part of our lives. Many people listen to music and we take, we participate in a lot of arts uh just generally so students **should** be able to to have that opportunity to uh to explore music and art. Uh it makes work complete and more eclectic mix in that courses and uh basically appreciation of art and music is central for proper academic development of any uh of any student, I mean uh, you don't only concentrate on one thing, you have a more rounded education uh. Also art and music have a long history, you be **able to** appreciate a culture, understand a lot of things in your community that have developed uh through art and music. They're very central to to culture and I think any student **should** be able to appreciate culture when they know mus-

Example J5 —Task 3 (Groundwater, Integrated), Level 3

-in the valley area are use, uh use to have #1 groundwa- water for irrigation but in the late eighteen hundreds it has been over pump. That started the land subsidence in the nineteen twenties. Uh, and then in the nineteen seventies uh, the amount of ground eh water drop drastically um, #1 which cause the water surface to si-, uh to sink the water surface about one hundred twenty meters and the ground sink about eight and a half meters. And #2 um, the efforts that were made to solve the problem is to reduce, reduce the pumping of the groundwater and then they start to import water. Uh but in #9 but uh in the nineteen nineties uh, they were forced to #2 um #1 pump the wa- groundwater again, again because the drought uh there were no rain for long time and that cause of the drop, #1 the drop of the whole surface of the entire valley.

Example J6 —Task 5 (Innate, Integrated), Level 4

Well, Robert Fantz wanted to know if infants could see a patterned and organized world #1 so he, he was showing some some face drawings to to children, #1 uh to see which one #2 did they pay more attention and he discovered that infant of all ages #1 preferred the the picture that was more realistic, #1 that was the the face of a real human and he confirmed that infants preferred the the realistic, the realistic things and that they saw a pattern in the world #1 and they **could** explore it. #4 Um that it.

Example J7 —Task 5 (Innate, Integrated), Level 4

The experiment is done to see if monkeys **can** count. Uh they are shown different images, uh, #1 each one with um, different uh #2 quantity of uh objects, like the first shows a circle, the second one's two squares and the third one three other objects and like that, from one through four. The monkeys then uh are or- they **have to** order them in the ascending order. First touch the- the image with one object then the one with two objects and if they get the right order they get a reward. Um, #1 this showed that they have uh, the concept of numbers, uh they have the cognitive ability of numeracy. They **can** even differentiate larger numbers from five to nine, uh without even a reward and seventy-five percent of them did it uh right. So this mean they understand this concept. Uh, and numeracy's a profound uh cognitive ability, not even one year babies can do it. So uh this show that um, monkeys have a, #2 the numeracy um capability without having the language so it means there's not a connection between numeracy and language, uh as a many psychologists thought before.

Test of English as a Foreign Language
PO Box 6155
Princeton, NJ 08541-6155
USA

To obtain more information about TOEFL programs and services, use one of the following:

Phone: 1-877-863-3546
(US, US Territories*, and Canada)

1-609-771-7100
(all other locations)

Email: toefl@ets.org

Web site: www.ets.org/toefl

* America Samoa, Guam, Puerto Rico, and US Virgin Islands

I.N. 728375